Journal of Experimental Psychology: Learning, Memory, and Cognition

The Role of Functional Form in Causal-Based Categorization

Bob Rehder

Online First Publication, August 11, 2014. http://dx.doi.org/10.1037/xlm0000048

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/14/$12.00 http://dx.doi.org/10.1037/xlm0000048

The Role of Functional Form in Causal-Based Categorization

Bob Rehder
New York University

Two experiments tested how the *functional form* of the causal relations that link features of categories affects category-based inferences. Whereas *independent causes* can each bring about an effect by themselves, *conjunctive causes* all need to be present for an effect to occur. The causal model view of category representations is extended to include a representation of conjunctive causes and then predictions are derived for 3 category-based judgments: classification, conditional feature predictions, and feature likelihoods. Experiment 1 revealed that subjects' judgments on all 3 tasks were not only sensitive to whether causes were independent or conjunctive but also conformed to the causal model predictions, albeit with an important exception. Experiment 2 revealed that inferences with independent and conjunctive causes were affected quite differently by a manipulation of the strengths of the causal relations (and in the manner predicted by the model). This is the 1st study to show how a single representation of a category's causal knowledge can account for 3 category-based judgments with the same model parameters. Other models of causal-based categories are unable to account for the observed effects.

Keywords: causal reasoning, causal-based categorization, interactive causes, conjunctive causes

Supplemental materials: http://dx.doi.org/10.1037/xlm0000048.supp

Categories are surprisingly complex and varied. Some categories seem to have a (relatively) simple structure. As a child, I learn to identify some parts of my toys as wheels and some letters in words as *t*. Accordingly, much of the field has devoted itself to the categorization of stimuli with a small number of perceptual dimensions, testing in the lab subjects' ability to learn to classify stimuli such as Gabor patches or rectangles that vary in height and width. Other categories, in contrast, have an internal structure that is imbued with other sorts of knowledge. Because of its ubiquity in our conceptual structures (Ahn, Marsh, Luhmann, & Lee, 2002; Kim & Ahn, 2002a, 2002b; Sloman, Love, & Ahn, 1998), one particular type of knowledge—the *causal relations* that obtain between features of categories—has received special attention. For example, even if you know little about cars, you probably have at least some vague notion that cars not only *have* (as features) gasoline, spark plugs, radiators, fans, and emit carbon monoxide but also that these features causally interact—that the spark plugs are somehow involved in the combustion of the gasoline, that the combustion produces carbon monoxide and heat, and that the radiator and fan somehow work to dissipate the latter.

Past research has focused on how the topology and strength of the links that make up a category's network of interfeature causal relations affect judgments of category membership (see Ahn & Kim, 2001, and Rehder, 2010, for reviews). Two important empirical effects have been established. First, the *causal status effect* is the phenomenon in which a feature is more important to category membership to the extent it is more "causal," that is, to the extent that it has many other features that it (directly or indirectly) causes. For example, consider the simple causal network in Figure 1A in which two category features, C_1 and C_2, are causes of a third, E. All else being equal, the causal status effect suggests that each of the causes carry greater weight on categorization decisions than E, because the former have one effect (E), whereas E itself has none (Ahn, Kim, Lassaline, & Dennis, 2000; Kim & Ahn, 2002a, 2002b; Sloman et al., 1998).[1]

A second type of empirical effect is that classifiers are sensitive to whether potential category members exhibit not only the right features considered individually but also the right feature *combinations*, a phenomenon known as the *coherence effect*. Referring again to Figure 1A, classifiers are not only sensitive to the weights of, say, C_1 and E, but also whether those features manifest the interfeature correlation one would expect on the basis of the causal relation (C_1 and E either both present or both absent). This effect obtains above and beyond the contribution of the C_1 and E considered individually (Ahn et al., 2002; Hayes & Rehder, 2012; Marsh & Ahn, 2006; Mayrhofer & Rothe, 2012; Rehder 2003a, 2003b; Rehder & Hastie, 2001; Rehder & Kim, 2006; see also Barrett, Abdi, Murphy, & Gallagher, 1993; Murphy & Wis-

This work was supported by Air Force Office of Scientific Research Grant FA9550-09-NL404.

Correspondence concerning this article should be addressed to Bob Rehder, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003. E-mail: bob.rehder@nyu.edu

[1] Subsequent research has uncovered a number of important qualifications to the causal status effect. First, because a feature's importance also increases as function of its number of causes, a feature with many causes can sometimes exceed that of the causes themselves (Rehder, 2003a; Rehder & Hastie, 2001; Rehder & Kim, 2006). Second, the magnitude of the causal status effect also varies with the strength of the causal relation: The difference in importance between a cause and effect decreases as the strength of the causal link increases, disappearing entirely (and sometimes reversing in direction) when the link is deterministic (Rehder & Kim, 2010). I return to both of these effects later in the article.

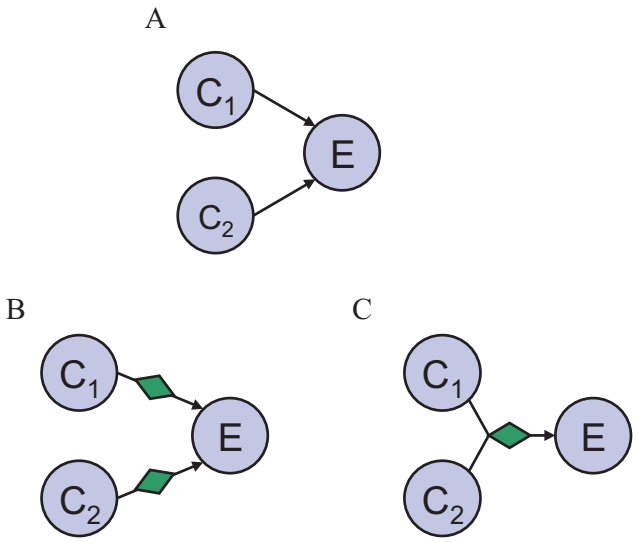

Figure 1. A: A simple causal network in which two variables cause a third. B and C: Two alternative interpretations of that network. See the online article for a color version of this figure.

niewski, 1989; Rehder & Ross, 2001; Wisniewski, 1995, for examples of coherence effects arising from interfeature relations that are not explicitly causal). As is the case with the causal status effect (see Footnote 1), the coherence effect is moderated by the strength of the causal relations, becoming larger as the links become stronger (Rehder & Kim, 2010).

This article tests the effect of a new sort of property of categories' causal networks, namely, the *functional form* of the relations between the causes and their effects. Returning again to the example in Figure 1A, there are many ways that C_1, C_2, and E might be related. C_1 and C_2 might be *independent* causes of E in the sense that they each bring about E via unrelated causal processes, as shown in Figure 1B. In the figure, diamonds represent independent generative causal mechanisms representing that E might be brought about by C_1 *or* C_2. In contrast, Figure 1C represents a situation in which C_1 and C_2 are *interactive* causes of E, that is, cases in which the causal influence of C_1 on E depends on the state of C_2, and vice versa. Note that this interaction might itself take on a number of forms. For example, the conjunction of two or more variables is often necessary for an outcome to occur. A spark may only produce fire if there is fuel to ignite; a virus may only cause disease if one's immune system is suppressed; the motive to commit murder may result in death only if the means to carry out the crime are available. In Figure 1C, if C_1 and C_2 are conjunctive causes of E, E is brought about only when C_1 *and* C_2 are present. When one of the conjuncts is consistently present, it is often construed as an *enabler* (Cheng & Novick, 1991). For example, the presence of oxygen enables fire given spark and fuel. In contrast, a *disabler* interacts with an existing cause by preventing its normal outcome (Park & Sloman, 2013; Walsh & Sloman, 2004, 2008).

Research from other domains provides ample reason to suspect that the sort of differences in functional form depicted in Figure 1 will be included in people's representations of categories. First, people learning new causal relations are sensitive to how the influences of multiple causes of a single effect are combined. Whereas people may assume that causes are independent by default (Cheng, 1997; Sobel, Tenenbaum, & Gopnik, 2004), prior knowledge and experience can lead people to assume a conjunctive integration rule instead (Lucas & Griffiths, 2010; Novick & Cheng, 2004; see also Kemp, Goodman, & Tenenbaum, 2010). When causes are continuous variables, prior knowledge helps determine whether people add or average them in order to predict a common effect (Waldmann, 2007; Zelazo & Shultz, 1989). Second, conditional reasoning research has shown how argument acceptability is affected by whether a conditional (if–then) statement is interpreted as reflecting a causal relation, but one with disabling conditions. For example, reasoners are less likely to accept the validity of a modus ponens argument when they can think of counterexamples reflecting the operation of disablers (Byrne, Espino, & Santamaria, 1999; Cummins, 1995; de Neys, Schaeken, d'Ydewalle, 2003a, 2003b, 2005; Fernbach & Erb, 2013; Frosch & Byrne, 2012; Geiger & Oberauer, 2007; Markovits & Potvin, 2001; Thompson, 1995).

The goal of this article was to test whether people's category-based inferences are sensitive to different functional forms between causes and effects. In particular, the following experiments test whether classifiers honor the distinction between independent and conjunctive causes. To this end, the following section specifies the (independent or conjunctive) generative functions that relate an effect and its causes and derives the different patterns of classification implied by those functions, predictions that are then tested in the following experiments.

Establishing that categorizers are sensitive to the functional form linking causes and effects will have implications for models that have been offered as accounts of how causal knowledge about categories is represented. For example, according to Sloman et al.'s (1998) *dependency model*, features are more important to category membership (i.e., are more *conceptually central*) as a function of the number and strength of the relations they have with *dependents*, that is, other features that depend on them (directly or indirectly). A causal relation is an example of a dependency relation in which the effect depends on its cause. Second, according to Rehder and colleagues' *generative model* (Rehder, 2003a, 2003b; Rehder & Kim, 2006), interfeature causal relations are represented as probabilistic causal mechanisms, and objects likely to have been generated by a category's causal model are considered to be good category members and those unlikely to be generated are poor ones. The generative model builds on the popular framework for modeling learning and reasoning with causal knowledge known as *Bayesian networks* or *causal graphical models* (hereafter, CGMs). Because CGMs specify the patterns of dependence and independence between variables but not the functional form of the cause–effect relations, they provide a natural framework for representing independent and conjunctive causes. In contrast, because it specifies that classification only depends on the features' dependency structure, the dependency model does not naturally predict sensitivity to differences in functional form. For example, the independent and conjunctive causes in Figures 1B and 1C both reduce to the same dependency structure, namely, the one in Figure 1A. I consider

possible extensions to the dependency model after the experiments have been reported.

The article's second goal was to bolster claims regarding the representation of categories' causal knowledge by showing that those representations mediate performance on multiple tasks. Whereas past model comparisons in this domain have focused on category membership judgments, any theory that purports to characterize categorical representations should also account for the many other sorts of category-based judgments. Accordingly, the section below derives how the proposed causal model representation of independent and conjunctive causes applies to not only classification but also two other tasks as well, namely, conditional inferences and judgments of features' likelihood within the category. Note that by showing that they account for performance on multiple tasks, the experiments will provide the converging evidence necessary to help establish the psychological reality of the proposed representations. To my knowledge, no previous tests of categorical knowledge have used three different sorts of judgments in a single study.

A final goal was to assess proposals regarding how people often augment category-based representations with additional knowledge. For example, Rehder and Burnett (2005) tested people's causal inferences by instructing them on categories whose features were causally related in a number of different topologies, including an independent cause network similar to the one in Figure 1B. Although subjects' inferences honored many of the predictions of CGMs, they also exhibited systematic deviations from those predictions, deviations that Rehder and Burnett argued were consistent with subjects assuming that there were causal links among features in addition to those provided by the experimenter. The following experiments test this claim by assessing whether the influence of this additional knowledge generalizes to a wider set of conditions. First, the same type of deviations observed for inferences with an independent cause network should appear for a conjunctive cause network. Second, that knowledge should influence not only causal inferences but judgments of categorization and feature likelihood as well. Accordingly, the following sections describe how this additional knowledge should affect all three types of judgments for both kinds of causal networks.

Category-Based Reasoning With Independent and Conjunctive Causes

To assess how people reason with independent and conjunctive causes, subjects were instructed on a novel category with six features. For example, subjects who learned about Romanian Rogos (a type of automobile) were told that Rogos have a number of typical or characteristic features (e.g., butane-laden fuel, a loose fuel filter gasket, hot engine temperature, etc.). In addition, subjects were instructed on the interfeature causal relations shown in Figure 2A. Features IC_1 and IC_2 were described as independent causes of IE, whereas CC_1 and CC_2 were described as conjunctive causes of CE. Subjects were then presented with a series of test questions for which they categorized and made conditional probability and feature likelihood judgments.

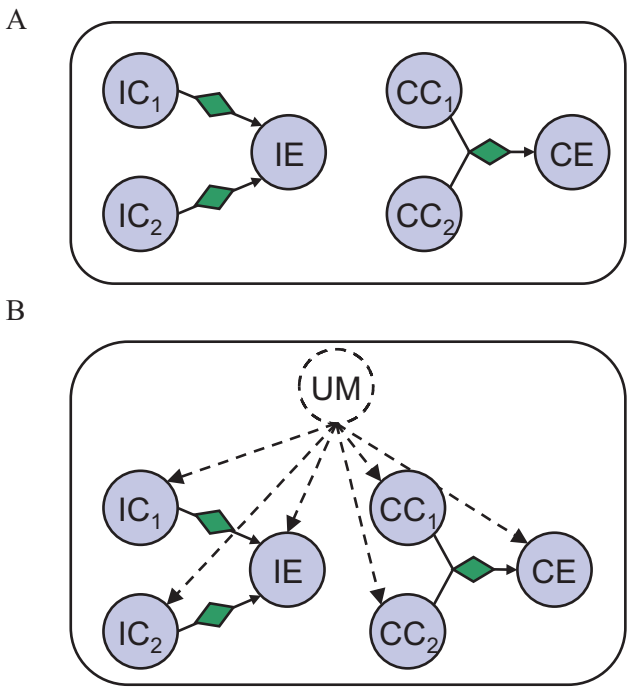

Figure 2. A: A schematic representation of the category features and causal relations learned in Experiment 1. B: The same model augmented with an underlying mechanism (UM) that is causally related to all variables. IC = independent cause; IE = independent effect; CC = conjunctive cause; CE = conjunctive effect. See the online article for a color version of this figure.

To derive predictions for this experiment, I first specify the joint probability distribution for each of the two CGMs represented by the subnetworks in Figure 2A and then use those distributions to derive expected inferences. The joint distribution for the independent cause network, $p_k(IC_1, IC_2, IE)$, is the probability that IC_1, IC_2, and IE will take any particular combination of values in category k. From the axioms of probability theory, it follows that,

$$p_k(IC_1, IC_2, IE) = p_k(IE \mid IC_1, IC_2)\, p_k(IC_1, IC_2). \quad (1)$$

Because IC_1 and IC_2 have no common causes (and because the *completeness* constraint associated with CGMs stipulates that they have no common causes that are hidden, i.e., that are not included in Figure 3A, i.e., Spirtes, Glymour, & Scheines, 2000), they can be assumed to be independent. Equation 1 thus becomes,

$$p_k(IC_1,\ IC_2,\ IE) = p_k(IE \mid IC_1, IC_2) p_k(IC_1) p_k(IC_2). \quad (2)$$

$p_k(IE \mid IC_1, IC_2)$ can be written as a function of parameters that characterize the generative causal mechanisms that relate IE to its causes. Specifically, let $m_{IC_1,IE}$ and $m_{IC_2,IE}$ represent the probabilities that those mechanisms will produce IE when IC_1 and IC_2 are present, respectively. In terms introduced by Cheng (1997), these probabilities refer to the "power" of the causes. In addition, to allow for the possibility that IE has additional causes not shown in Figure 2A, let b_{IE} represent the probability that IE will be brought about by one or more hidden causes.

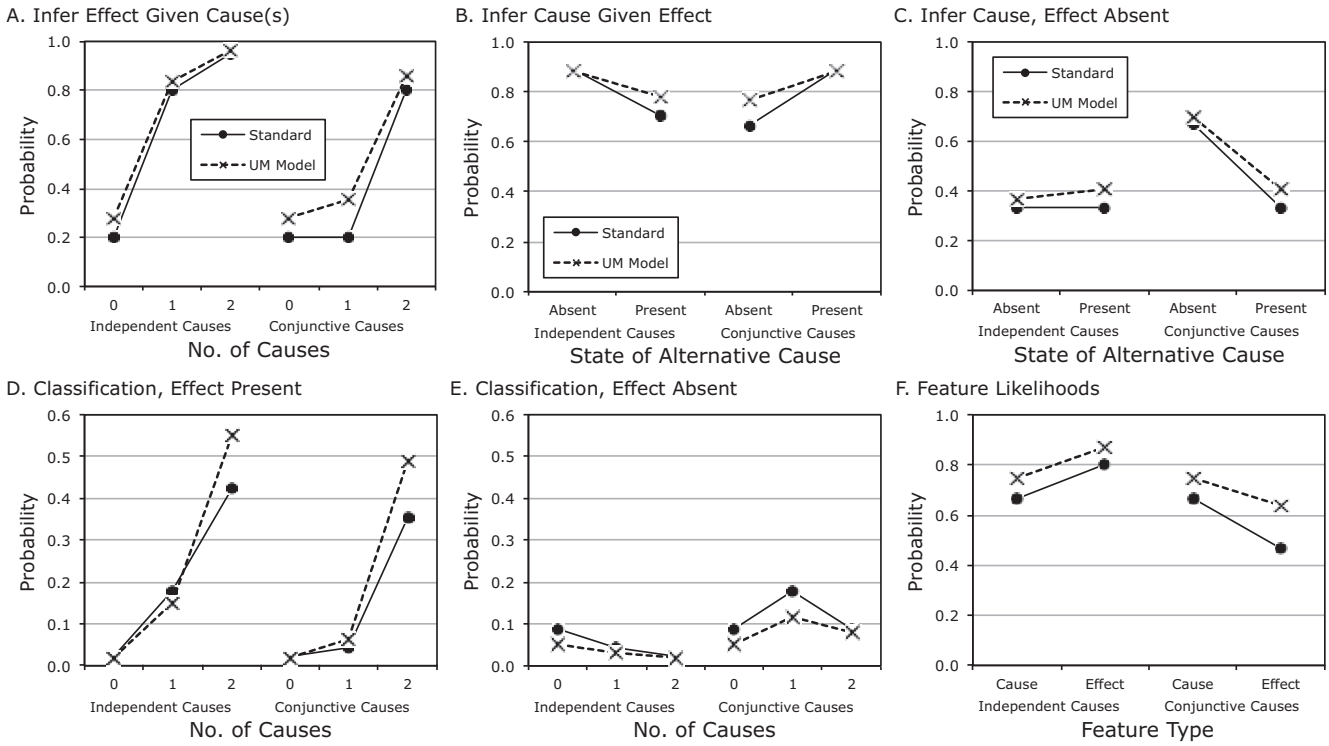

Figure 3. Predictions for the independent and conjunctive cause networks. Solid lines represent are the predictions for the standard causal graphical models in Figure 2A, and dashed lines are those for the underlying mechanism (UM) model in Figure 2B. Panels A–C present the predicted probability that a feature will be present as a function of the presence or absence of other features. A: The probability of the effect as a function of the number of causes present. B: The probability of a cause as a function of whether the other cause is present, assuming the effect is present. C: The probability of a cause as a function of whether the other cause is present, assuming the effect is absent. Panels D and E present the predicted probability that a configuration of features will be generated by the category (and thus the probability that an object with those features is a member of the category). D: The probability of each feature configuration in which the effect is present. E: The probability of each configuration in which the effect is absent. F: The within-category marginal probability of the individual cause and effect features.

Given these definitions, the probability that IE is present is given by the familiar noisy-or equation,

$$p_k(IE = 1 \,|\, IC_1, IC_2) = 1 - (1 - b_{IE}) \prod_{i=1,2} (1 - m_{IC_i,IE})^{ind(IC_i)}, \tag{3}$$

where $ind(IC_i)$ returns 1 when IC_i is present and 0 otherwise. For example, the probability that IE is present when IC_1 is present and IC_2 is absent is,

$$p_k(IE = 1 \,|\, IC_1 = 1, IC_2 = 0) = 1 - (1 - b_{IE})(1 - m_{IC_1,IE})^1 \times (1 - m_{IC_2,IE})^0 = m_{IC_1,IE} + b_{IE} - m_{IC_1,IE} m_{IC_1,IE}.$$

That is, IE is brought about by (the causal mechanism associated with) IC_1 *or* by some hidden causes (of net strength b_{IE}).

Equations 2 and 3 are sufficient to specify the probability of any combination of IC_1, IC_2, and IE. These expressions are shown in the top half of Table 1. For example, the probability that IC_1 and IE are present and IC_2 absent is,

$$p_k(IC_1 = 1, IC_2 = 0, IE = 1) = [1 - (1 - b_{IE}) \times (1 - m_{IC_1,IE})] c_{IC_1} (1 - c_{IC_2}),$$

where c_{IC_1} and c_{IC_2} are the probabilities that IC_1 and IC_2, will appear in members of category k, respectively.

The joint distribution for the conjunctive cause network, $p_k(CC_1, CC_2, CE)$, can be written in a manner analogous to Equation 2,

$$p_k(CC_1, CC_2, CE) = p_k(CE \,|\, CC_1, CC_2)\, p_k(CC_1) p_k(CC_2). \tag{4}$$

The conjunctive cause network in Figure 2A differs from independent causes in having one generative causal mechanism. I extend the notion of causal power to conjunctive causes by assuming that when CC_1 and CC_2 are both present, that mechanism will bring about CE with probability $m_{CC_1,CC_2,CE}$. Thus,

$$p_k(CE = 1 \,|\, CC_1, CC_2) = 1 - (1 - b_{CE}) \times (1 - m_{CC_1,CC_2,CE})^{ind(CC_1) ind(CC_2)}, \tag{5}$$

Table 1
Joint Probability Distributions for Independent and Conjunctive Causes

Independent causes				$c_{IC_1} = c_{IC_2} = .67$; $m_{IC_1,IE} = m_{IC_2,IE} = .75$; $b_{IE} = .20$
IC_1	IC_2	IE	$p_k(IC_1, IC_2, IE)$	
1	1	1	$[1 - (1 - m_{IC_1,IE})(1 - m_{IC_2,IE})(1 - b_{IE})]c_{IC_1}c_{IC_2}$	.422
1	1	0	$(1 - m_{IC_1,IE})(1 - m_{IC_2,IE})(1 - b_{IE})c_{IC_1}c_{IC_2}$	.022
1	0	1	$[1 - (1 - m_{IC_1,IE})(1 - b_{IE})]c_{IC_1}(1 - c_{IC_2})$	.178
0	1	1	$[1 - (1 - m_{IC_2,IE})(1 - b_{IE})](1 - c_{IC_1})c_{IC_2}$	.178
0	0	1	$b_{IE}(1 - c_{IC_1})(1 - c_{IC_2})$	.022
0	1	0	$(1 - m_{IC_2,IE})(1 - b_{IE})(1 - c_{IC_1})c_{IC_2}$	.044
1	0	0	$(1 - m_{IC_1,IE})(1 - b_{IE})c_{IC_1}(1 - c_{IC_2})$	.044
0	0	0	$(1 - b_{IE})(1 - c_{IC_1})(1 - c_{IC_2})$	.089
Conjunctive causes				$c_{CC_1} = c_{CC_2} = .67$; $m_{CC_1,CC_2,CE} = .75$; $b_{CE} = .20$
CC_1	CC_2	CE	$p_k(CC_1, CC_2, CE)$	
1	1	1	$[1 - (1 - m_{CC_1,CC_2,CE})(1 - b_{CE})]c_{CC_1}c_{CC_2}$	.356
1	1	0	$(1 - m_{CC_1,CC_2,CE})(1 - b_{CE})c_{CC_1}c_{CC_2}$	.089
1	0	1	$b_{CE}c_{CC_1}(1 - c_{CC_2})$	.044
0	1	1	$b_{CE}(1 - c_{CC_1})c_{CC_2}$	.044
0	0	1	$b_{CE}(1 - c_{CC_1})(1 - c_{CC_2})$	.022
0	1	0	$(1 - b_{CE})(1 - c_{CC_1})c_{CC_2}$	.178
1	0	0	$(1 - b_{CE})c_{CC_1}(1 - c_{CC_2})$	.178
0	0	0	$(1 - b_{CE})(1 - c_{CC_1})(1 - c_{CC_2})$	.089

Note. IC = independent cause; IE = independent effect; CC = conjunctive cause; CE = conjunctive effect.

where b_{CE} is the probability that CE will be brought about by other (unidentified) cause(s). For example, the probability that CE is present when CC_1 is present and CC_2 is absent is,

$$p_k(CE = 1 \mid CC_1 = 1, CC_2 = 0) = 1 - (1 - b_{CE}) \times (1 - m_{CC_1,CC_2,CE})^{(1)(0)} = b_{CE}.$$

Equations 4 and 5 are sufficient to specify the probability of any combination of CC_1, CC_2, and CE, as shown in the bottom half of Table 1. c_{CC_1} and c_{CC_2} are the probabilities that CC_1 and CC_2 will appear in members of category k, respectively. Note that whereas Equation 5 defines a conjunctive causal mechanism as one that is only effective when both causes are present, Novick and Cheng's (2004) framework for learning interactive causes is more general in that it allows each of the causes to also have an independent influence on the outcome. Analogously, the animal learning literature includes models in which cues can have both an "elemental" (i.e., independent) and configural (interactive) effect on the outcome (Gluck & Bower, 1988; Pearce, 1987, 1994, 2002; Wagner & Rescorla, 1972). In fact, it is straightforward to generalize Equation 5 to the case in which two causes may have both independent and conjunctive influences on the effect,

$$p_k(CE = 1 \mid CC_1, CC_2) = 1 - (1 - b_{CE}) \times (1 - m_{CC_1,CC_2,CE})^{ind(CC_1)ind(CC_2)} \prod_{i=1,2} (1 - m_{CC_i,CE})^{ind(CC_i)}, \quad (6)$$

where $m_{CC_1,CE}$ and $m_{CC_2,CE}$ represent the independent influence that CC_1 and CC_2 have on CE. Nevertheless, the materials on which subjects are instructed in the upcoming experiment will emphasize that CC_1 and CC_2 have only a conjunctive influence on CE, and so this article thus uses the more specific definition of conjunctive causes represented by Equation 5.

Given the joint distributions in Table 1, it is straightforward to compute predictions for the three types of judgments assessed in this study, namely, judgments of category membership, conditional probability (of the presence of a feature given other features), and the likelihood of individual features.

Categorization Judgments

Subjects completed a classification task in which they were presented with objects with lists of features and asked whether they were likely category members. For example, subjects were presented with an automobile with a number of features (e.g., butane-laden fuel, a normal fuel filter gasket, hot engine temperature, etc.) and asked to judge the likelihood that it was a Rogo. To derive classification predictions, the assumptions of the generative model are adopted; that is, an object's degree of category membership is proportional to the probability that its features were generated by the category's causal network (Rehder, 2003a, 2003b; Rehder & Kim, 2006, 2010). These probabilities of course are just those given by the joint distributions

in Table 1 that specifies the probability of any combination of the presence/absence of each feature.[2]

To demonstrate the qualitative pattern of category membership judgments supported by independent and conjunctive causes, the joint distributions were instantiated with parameter values that are hypothetical but reasonable in light of conditions established in the upcoming experiment. Because they are described as typical category features, each cause is assumed to be moderately prevalent among category members (the *c*s = .67), each causal mechanism is moderately strong (the *m*s = .75), and the alternative causes of the effect features are weak (the *b*s = .20). The classification predictions generated from these parameter values are presented in Figures 3D and 3E as those of the "standard model." Figure 3 also includes predictions for an underlying mechanism (UM) model discussed below.

Figure 3D presents the probability of objects in which the effect is present as a function of the number of causes and type of network. For both independent and conjunctive causes, when the effect is present, the most likely objects of course are those in which both causes are also present and the least likely are those in which both causes are absent. However, whereas an object in which one cause and the effect are present is moderately probable for independent causes, it is very improbable for conjunctive causes, an example of a coherence effect.

Figure 3E presents the probability of objects in which the effect is absent. For independent causes, an object becomes less probable as it has more causes, a result that obtains because the presence of causes is inconsistent with an absent effect. Note that this prediction reveals that, as a result of coherence, an object with more typical features can be a worse category member if those features violate the category's causal relationships. A different pattern emerges for conjunctive causes. Because two causes are needed to produce the effect, an object missing an effect is incoherent only when both causes are present. As a result, an object in which one cause is present and the effect is absent is moderately probable for conjunctive causes but very improbable for independent causes.

In summary, these predictions indicate that very different patterns of coherence should emerge during the classification task for independent and conjunctive causes. An item with the effect and one cause is a likely category member for independent links but not conjunctive ones, and the reverse pattern holds for items without the effect and one cause. The following experiments test for the interaction between network type and the quadratic effect of the number of causes shown in Figures 3D and 3E.

Conditional Feature Inferences

Subjects also completed a feature prediction task in which they were presented with objects with a partial list of features and asked to predict the presence of another feature. For example, subjects were presented with a Rogo with butane-laden fuel and asked to predict whether it has a hot engine temperature. To demonstrate the qualitative pattern of conditional probability judgments supported by independent and conjunctive causes, the joint distributions were instantiated with same parameter values as in Figures 3D and 3E, from which any conditional inference can be derived. For example, the probability of IC_2 conditioned on the presence of IE, and the absence of IC_1 is,

$$p_k(IC_2 = 1 \mid IC_1 = 0,\ IE = 1) = p_k(IC_2 = 1, IC_1 = 0,\ IE = 1)/$$

$$p_k(IC_1 = 0,\ IE = 1) = .178/(.178 + .022) = .890.$$

Predictions for the independent and conjunctive cause networks for three distinct types of inference problems are shown in Figures 3A–C. First, Figure 3A presents the probability of the effect as a function of the number of causes that are present for both the independent and conjunctive cause networks. For independent causes, the probability of the effect of course increases monotonically with the number of causes. (The probability of the effect is .20 even when both causes are absent because of the potential of additional causes, represented by $b_{IE} = .20$.) In contrast, for conjunctive causes, the probability of the effect increases from its baseline of .20 only when *both* causes are present. That is, as was the case in Figures 3D and 3E, an interaction between network type and the quadratic effect of the number of causes should obtain.

Figure 3B presents inferences to a cause when the effect is present as a function of the state of the other cause. For an independent cause network, the probability of a cause is higher when the other cause is absent versus present. This is the well-known *explaining away* phenomenon in which the presence of one cause that accounts for an effect makes other causes less likely (Pearl, 1988, 2000; Spirtes et al., 2000).[3] For example, although you may initially suspect that a burglar is the reason your house alarm has sounded, the presence of a potential alternative cause (an earthquake) will reduce your confidence that a burglary is underway. Morris and Larrick (1995) have shown how explaining away is expected under a wide range of conditions (see also Jones & Harris, 1967; McClure, 1998).

The conjunctive cause network, in contrast, shows the opposite pattern, namely, the probability of a cause is *lower* when the other

[2] In contexts in which other categories are defined, a classifier's judgment that an object *o* belongs to category *k* of course should be relative to the probability that it belongs to one of the other categories, that is, according to Bayes' rule,

$$p(o_C = k \mid o_F) = \frac{p_k(o_F)p(o_c = k)}{\sum_{j \in K} p_j(o_F)p(o_c = j)},$$

where o_C is *o*'s category label, o_F are its features, and *k* is the set of potential categories. Past research (Ahn et al., 2000; Rehder, 2003a, 2003b; Rehder & Kim, 2006, 2010; Sloman et al., 1998) suggests that when category membership ratings are made with respect to a single, experimentally defined category, subjects simply generate $p_k(o_F)$. Alternatively, those judgments may be construed as being made in accord with Bayes' rule,

$$p(o_C = k \mid o_F) = \frac{p_k(o_F)p(o_c = k)}{p_k(o_F)p(o_c = k) + A},$$

where $A = \sum_{j \in K, j \neq k} p_j(o_F)p(o_c = j)$ is a constant representing the probability that o_F was generated by the one of the members of an (unspecified) set of alternative categories. $p(o_C = k \mid o_F)$ defined in this manner is monotonically related to $p_k(o_F)$.

[3] Note that the predictions in Figure 3B compare the probability of cause given the presence or absence of the alternative cause, e.g., $p_k(IC_2 = 1 \mid IC_1 = 1, IE = 1) < p_k(IC_2 = 1 \mid IC_1 = 0, IE = 1)$. Explaining away also refers to the fact that $p_k(IC_2 = 1 \mid IC_1 = 1, IE = 1)$ is generally lower than when no knowledge of the alternative cause is given, i.e., $p_k(IC_2 = 1 \mid IE = 1)$.

cause is absent versus present. For example, discovering that a murder suspect did not possess the means to carry out the crime (e.g., proximity to the victim) decreases his or her likely guilt (and leads the police to conclude that it was committed by some other, still unidentified suspect). I refer to this as the *exoneration effect*. This implication of treating two causes as conjunctive rather than independent has not been emphasized by previous investigators (cf. Kelley, 1972; Morris & Larrick, 1995).

Finally, Figure 3C presents the probability of a cause when the effect is absent as a function of the state of the other cause. On one hand, although independent causes are negatively correlated when the effect is present (the explaining away effect), they are uncorrelated when the effect is absent (and thus the probability of a cause is unaffected by the state of the other cause). In contrast, the probability of a conjunctive cause is lower when the other cause is present. This represents another form of exoneration: Your sibling, who promised to attend your Thanksgiving dinner but failed to arrive, is exonerated from responsibility when you learn that his or her airplane flight (the means needed to achieve the ends) was canceled due to a snowstorm.

Feature Likelihood Judgments

A third judgment type produced by subjects was a feature likelihood rating task in which each category feature was presented in isolation and subjects were asked to rate what proportion of category members have that feature. For example, subjects were presented with the feature butane-laden fuel and asked what percentage of Rogos had that feature. The proposed representations of independent and conjunctive causal networks make distinct predictions regarding the prevalence of the causes and effects, that is, their within-category marginal probabilities. Assuming that the marginal probabilities of the cause features (the *c* parameters), the causal strengths (the *m* parameters), and the background causes (the *b* parameters) are equated across the two networks, the central prediction is that the marginal probability of the independent effect IE will be greater than that of the conjunctive effect CE. This is the case because CE requires both of its causes to be present, whereas IE only requires one. Expressions that compute the within-category marginal probability of IE and CE are provided by Equations 7 and 8. These expressions are instantiated with the same parameter values as in Figures 3A–E.

$$p_k(IE = 1) = 1 - (1 - b_{IE})(1 - c_{IC_1} m_{IC_1,IE})(1 - c_{IC_2} m_{IC_2,IE})$$

$$= 1 - (1 - .20)(1 - (.67)(.75))(1 - (.67)(.75)) = .800 \quad (7)$$

$$p_k(CE = 1) = 1 - (1 - b_{CE})(1 - c_{CC_1} c_{CC_2} m_{CC_1,CC_2,CE})$$

$$= 1 - (1 - .20)(1 - (.67)(.67)(.75)) = .467. \quad (8)$$

These predictions, which are also presented in Figure 3F, illustrate an interaction in which the marginal probability of the independent effect IE relative to its causes IC_1 and IC_2 is greater than the marginal probability of the conjunctive effect CE relative to its causes CC_1 and CC_2.

The Typicality Effect and UM Hypothesis

The predictions presented in Figure 3 embody assumptions regarding the extent to which the network in Figure 2A is a complete depiction of the causal representations used by a reasoner. On one hand, a CGM need not be complete in the sense that variables may have exogenous influences (i.e., hidden causes) that are not part of the model; the strength of those influences that are captured by the *b* parameters specified earlier. Importantly, however, these influences are constrained to be uncorrelated across variables. This property, referred to as *causal sufficiency* (Spirites et al., 2000), is crucial because it enables claims regarding the assumptions of conditional independence among variables; in particular, it enables the causal Markov condition that specifies conditions under which variables are conditionally dependent or independent depending on the state of other variables (Pearl, 1988, 2000; Spirtes et al., 2000).

As mentioned, Rehder and Burnett's (2005) test of how people infer causally related category features found systematic deviations from the predictions of CGMs. The pattern of the results was the same regardless of network topology: Subjects rated a target feature as more likely as a function of the number of other category features already present, even when the Markov condition stipulated that those features were conditionally independent of the target. Rehder and Burnett referred to this result as a *typicality effect* in which the presence of typical features implies the presence of still more typical features and proposed that it arises because people reason with additional causal knowledge—in particular, that people assume that category features are related via a hidden common cause. The assumption of a common underlying causal mechanism (UM) can be viewed as summarizing the vague beliefs that people have about the hidden causal processes associated with categories (Gelman, 2003; Medin & Ortony, 1989).

This article assesses whether people's causal inferences conform to generative representations of independent and conjunctive causes augmented with an assumption of underlying causal mechanism. For example, Figure 2B presents the network that results if the category features in Figure 2A are each related to a single UM. Generating a quantitative example of how this network yields typicality effects requires specifying two additional parameters: c_{UM}, the probability that UM is present, and m_{UM}, the power of the causal links between it and the six category features. The UM's causal influence on a category feature is assumed to combine with its other influences according to a noisy-or function (see the Appendix for details). The predictions of the UM model assuming the same parameters as the standard causal model in Figure 2A (*c*s = .67, *m*s = .75, and *b*s = .20) and also assuming that $c_{UM} = m_{UM} = .50$ are shown in Figure 3 superimposed on those of the standard model. For the independent cause subnetwork, the cause features are no longer conditionally independent when the state of the effect is unknown (see the left side of Figure 3C), because the presence of one cause suggests the likely presence of UM, which in turn makes the presence of the other cause more likely. For the conjunctive cause subnetwork, the effect is no longer conditionally independent of a cause when the other cause is absent (see the right side of Figure 3A), because one can reason from the cause to the UM to the effect. These effects arise because the UM provides an alternative inferential path such that the presence of one typical feature increases the likelihood of other typical features. Later I present fits of the UM model to subjects' ratings, testing whether it accounts for judgments of category membership and feature likelihoods as well.

Experiment 1

Experiment 1 assesses whether people's judgments are consistent with the predictions for inference, classification, and feature likelihood tasks. As this is the first test of category-based reasoning with conjunctive causes, the initial goal was to test whether subjects manifest the qualitative phenomena that distinguish them from independent causes. Accordingly, subjects were not provided with values corresponding to the causal model parameters, that is, exact information about the probability of each cause (the *c* parameters), the strength of the causal links (the *m* parameters), or the possibility of alternative causes (the *b* parameters). Instead, I assess whether subjects exhibit, for example, explaining away and exoneration in judgments of conditional probability and the patterns of coherence in judgments of category membership associated with independent and conjunctive causes.

Method

Materials. Six novel categories were tested: two biological kinds (Kehoe Ants, Lake Victoria Shrimp), two nonliving natural kinds (Myastars [a type of star], Meteoric Sodium Carbonate), and two artifacts (Romanian Rogos, Neptune Personal Computers). Each category had six binary feature dimensions. One value on each dimension was described as typical of the category. For example, subjects who learned Romanian Rogos were told that "Most Rogos have a hot engine temperature, whereas some have a normal engine temperature"; "Most Rogos have loose fuel filter gasket, whereas some have a normal gasket"; and so on.

Subjects were also presented with three paragraphs describing the three causal relations in Figure 2A. The first sentence of each independent cause paragraph stated that one feature caused another. The first sentence of the conjunctive cause paragraph stated that two features "together" caused another. The rest of each paragraph described the mechanism responsible for the causal relationship. Table 2 presents an example of independent and conjunctive causes for Rogos. The full set of features and causal relationships for all six categories are presented in the supplementary material (see Tables S1–S6).

The assignment of the six typical category features to the causal roles in Figure 1 (IC_1, IC_2, IE, CC_1, CC_2, and CE) was balanced over subjects such that for each category, one triple of features played the role of IC_1, IC_2, and IE and the other played the role of CC_1, CC_2, and CE for half the subjects, and this assignment was reversed for the other half.

Procedure. Subjects first studied several computer screens of information about the category. Three initial screens presented the category's cover story and which features occurred in "most" versus "some" category members. The fourth screen presented the three paragraphs of causal information. When initially describing this screen, the experimenter read the first sentence of each paragraph, emphasizing that each independent cause produced the effect "by itself," whereas the two conjunctive causes "together" produced the effect.

A fifth screen presented a diagram like Figure 2A (with the names of the category's actual features). Just as in this figure, the independent cause diagram consisted of two arrows (one from each cause to the effect), whereas the conjunctive cause diagram consisted of two lines from the causes that joined into a single arrow leading to the effect. In the initial presentation of these screens, the experimenter reiterated how each independent cause produced the effect "by itself," whereas the conjunctive causes "together" produced the effect.

When ready, subjects took a multiple-choice test that tested them on this knowledge. While taking the test, subjects were free to return to the information screens; however, doing so obligated them to retake the test. The only way to proceed was to take the test all the way through without errors and without asking for help.

Subjects were then presented with inference and classification tests, balanced for order; the feature likelihood test was always presented last. During the inference test, subjects were presented with a total of 24 inference problems, 12 for each subnetwork. They were asked to (a) predict the effect given all possible states of the causes (four problems) and (b) predict each cause given all possible states of the effect and the other cause (eight problems). For example, subjects who learned Rogos were asked to suppose that a Rogo had been found that had butane-laden fuel and a loose fuel filter gasket and to judge how likely it was that it also had a hot engine temperature. Responses were entered by positioning a slider on a scale where the left end was labeled *Sure that it doesn't* and the right end was labeled *Sure that it does*. The position of the slider was recorded as a number in the range 0–100. The order of presentation of the 24 test items was randomized for each subject. So that judgments did not depend on subjects' ability to remember the causal relations, they were provided with a printed diagram

Table 2
Example Materials From One of the Counterbalancing Conditions Associated With the Romanian Rogos Experimental Category

Features	Causal relationships
High amounts of carbon monoxide [IC_1]	High amounts of carbon monoxide in the exhaust causes a long-lived generator. The carbon monoxide increases the pressure of the exhaust that enters the turbocharger, resulting in the turbocharger drawing less electricity from the generator, extending its life. [$IC_1 \rightarrow IE$]
Damaged fan belt [IC_2]	A damaged fan belt causes a long-lived generator. When the damaged fan belt slips, the generator turns at lower RPMs, which means that it lasts longer. [$IC_2 \rightarrow IE$]
Long-lived generator [IE]	Butane-laden fuel and loose fuel filter gaskets together cause a hot engine temperature. Loose fuel filter gaskets allow a small amount of fuel to seep into the engine bearings. This normally has no effect. However, if there is butane in the fuel, it undergoes a chemical reaction that creates heat as a by-product. Thus, when a car has both butane-laden fuel and a loose filter gasket, the engine runs at a hot temperature. [(CC_1,CC_2) $\rightarrow$ CE]
Butane-laden fuel [CC_1]	
Loose fuel filter gaskets [CC_2]	
Hot engine temperature [CE]	

Note. This condition's assignment of features and causal relationships to the causal roles shown in Figure 2A appear in brackets. IC = independent cause; IE = independent effect; CC = conjunctive cause; CE = conjunctive effect.

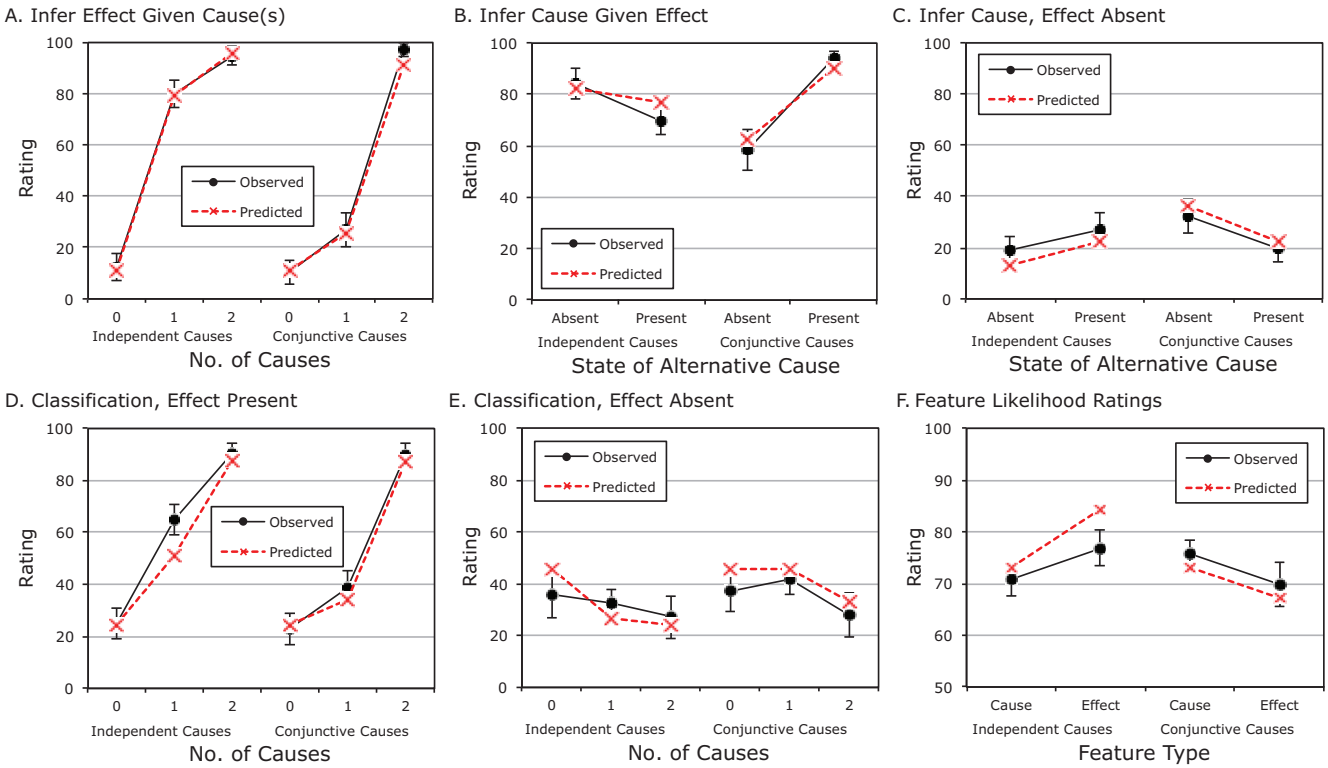

Figure 4. Empirical results from Experiment 1. Inference results are in Panels A–C, classification results are in Panels D–E, and feature likelihood results are in Panel F. Error bars are 95% confidence intervals. Fit of the underlying mechanism model (see Figure 2B) are superimposed on the observed data in each panel. See the online article for a color version of this figure.

similar to the one in Figure 2A during all three tests. Subjects were asked to make use of those causal relations in answering the questions.

Note that several of the 24 inference problems presented to subjects are logically equivalent in that they ask questions that should receive the same answer on the basis of the symmetrical causal information provided in the cover stores. For example, in Figure 2A, $p_k(IC_1 = 1 \mid IC_2 = 1, IE = 1)$ should equal $p_k(IC_2 = 1 \mid IC_1 = 1, IE = 1)$. Any differences in these responses reflect response noise or effects due to the materials of little theoretical interest. Responses to such equivalent problems were thus collapsed into 14 logically distinct inference problems.

During the classification test, subjects rated the category membership of 16 test items. Each item displayed three features from either the independent or the conjunctive subnetwork; no values were specified for the features from the other subnetwork. For example, subjects were told that an automobile had been found that had butane-laden fuel, a normal fuel filter gasket, and a hot engine temperature, and asked whether it was a Rogo. For each subnetwork, the three binary variables were instantiated with all possible combinations of values ($2^3 = 8$ test items for each subnetwork). Responses were entered by positioning a slider on a scale where the right end was labeled *Sure that it is* (a category member) and the left end was labeled *Sure that it isn't*. The position of the slider was scaled into the range 0–100. The order of the trials was randomized for each subject. Responses to logically equivalent classification trials were collapsed into 12 distinct types.

During the feature likelihood rating task, each of the two features on the six binary dimensions were presented on the computer screen, and subjects rated what proportion of all category members possessed that feature. The order of these trials was randomized for each subject. Responses to logically equivalent trials were collapsed into four distinct types. Experimental sessions lasted approximately 45 min.

Subjects. Forty-eight New York University undergraduates received course credit for participating in this experiment. There were three between-subject factors: the two assignments of physical features to their causal roles, the two task presentation orders, and which of the six categories was learned. Subjects were randomly assigned to these $2 \times 2 \times 6 = 24$ between-participant cells subject to the constraint that an equal number appeared in each cell.

Results

Initial analyses revealed no effects of which category subjects learned, the assignment of features to causal roles (i.e., the independent or conjunctive subnetwork), or task presentation order, and so the results are presented collapsed over these factors.

Classification results. Subjects' classification ratings are presented in Figures 4D and 4E and were generally in accord with the model's predictions. First, when the effect was present (see Figure 4D), the ratings varied in the manner shown in Figure 3D. Whereas objects received low ratings when both causes were absent and

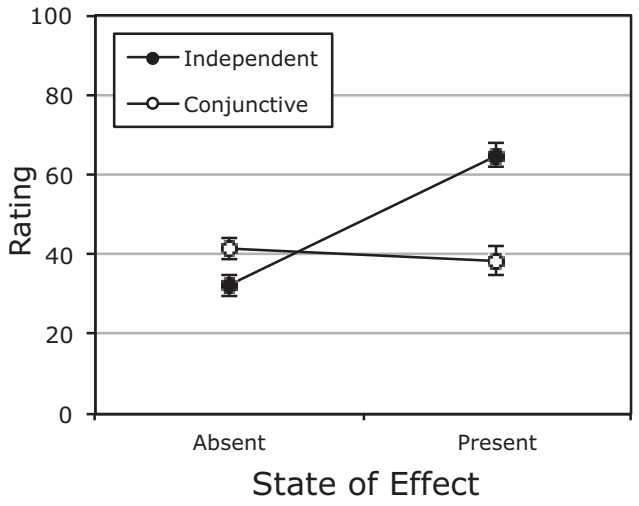

Figure 5. Classification results from Experiment 1 for items with one out of two causes present as a function of the state of the effect and network type. Error bars are standard errors.

received high ones when they were both present, objects received moderately high ratings when one cause was present when the causal relations were independent (64.8) but a much lower one when they were conjunctive (38.4). A 3 × 2 analysis of variance (ANOVA) of the data in Figure 4D revealed a main effect of the number of causes, $F(2, 94) = 236.35$, $MSE = 446$, $p < .0001$; a main effect of network type, $F(1, 47) = 25.14$, $MSE = 273$, $p < .0001$; and an interaction, $F(2, 94) = 17.42$, $MSE = 289$, $p < .0001$. The predicted interaction between network type and the quadratic trend of the number of causes, $F(1, 47) = 26.18$, $MSE = 767$, $p < .0001$, was significant, confirming that the two networks differed in how classification ratings increased with the number of causes. When only one cause was present, ratings were much higher for the independent versus conjunctive cause networks, $t(47) = 6.35$, $p < .0001$.

Second, when the effect was absent (see Figure 4E), categorization ratings decreased as the number of independent causes increased, reflecting the pattern of coherence effects anticipated by Figure 3E. For conjunctive causes, ratings reflected a nonmonotonic change in ratings as the number of causes increased, with the item with one cause receiving the highest ratings. An ANOVA of these data revealed marginal main effects of the number of causes and network type, $F(2, 94) = 2.40$, $MSE = 1{,}144$, $p = .11$; and, $F(1, 47) = 2.49$, $MSE = 430$, $p = .12$; and a marginal interaction, $F(1, 47) = 1.77$, $MSE = 279$, $p = .18$. Although the interaction between network type and the quadratic trend of the number of causes, $F(1, 47) = 3.13$, $MSE = 625$, $p = .08$, was also marginal; the more focused comparison on the item in which one cause was present revealed that, as predicted, ratings were higher for the conjunctive versus independent cause network (41.5 vs. 32.4), $t(47) = 2.26$, $p = .03$.

Figure 5 highlights the different patterns of coherence that emerged for independent versus conjunctive causes by presenting ratings for items with one out of two causes present. When the effect was present, ratings were much lower when the causes were conjunctive, suggesting that subjects reasoned that the presence of the conjunctive effect was incompatible with the presence of only one of its two causes. When the effect was absent, ratings were lower for independent causes, suggesting that they reasoned that the absence of the effect was incompatible with the presence of a single independent cause.

Feature inference results. Inference ratings are presented in Figures 4A–C and also generally reflected the qualitative predictions shown in Figures 3A–C. Starting with Figure 4A, subjects judged that the presence of the effect was rated to be very likely (ratings > 90) when two causes were present and very unlikely (< 15) when they were absent for both independent and conjunctive causes. When only one cause was present, ratings differed depending on the type of network. Subjects were much more likely to predict the effect when one cause was present when the causes were independent (rating of 79.9) as compared with conjunctive (26.8).

Figure 4A also reveals one way that the ratings for the conjunctive cause network differed from Figure 3A's predictions. Although the conjunctive effect should be equally probable regardless of whether zero or one cause is present, subjects judged that the effect was more probable in the presence of one cause (26.8) versus none (10.5). Of course, this result corresponds to the typicality effect described earlier in which the presence of a typical feature increases the likely presence of another typical feature, even if those features are conditionally independent.

Statistical analysis supported these conclusions. A 3 × 2 ANOVA of the data in Figure 4A revealed a main effect of the number of causes, $F(2, 94) = 470.36$, $MSE = 364$, $p < .0001$; a main effect of network type, $F(1, 47) = 65.91$, $MSE = 335$, $p < .0001$; and an interaction, $F(2, 94) = 82.62$, $MSE = 278$, $p < .0001$. As predicted, the more focused test of the interaction between network type and the quadratic trend of the number of causes, $F(1, 47) = 99.67$, $MSE = 917$, $p < .0001$, was significant, confirming that the two networks differed in how inference ratings increased with the number of causes. In particular, when only one cause was present, ratings were much higher for the independent versus conjunctive cause networks, $t(47) = 9.88$, $p < .0001$. Nevertheless, the conjunctive effect was rated more probable in the presence of one versus zero causes, $t(47) = 5.25$, $p < .0001$.

Figure 4B shows inference ratings when subjects predicted a cause when the effect is present. When causes were independent, subjects exhibited explaining away: The cause was rated higher when the other cause was absent (84.2) versus present (69.6). In contrast, this pattern was reversed for conjunctive causes (58.4 vs. 94.0), that is, subjects exonerated. A 2 × 2 ANOVA of the data in Figure 4B revealed an effect of the state of the other cause, $F(1, 47) = 9.68$, $MSE = 549$, $p < .01$; no effect of network type ($F < 1$); but an interaction, $F(1, 47) = 83.62$, $MSE = 362$, $p < .0001$. Ratings were higher when the other cause was absent, $t(47) = 3.03$, $p < .01$, when causes were independent (explaining away), whereas they were lower when they were conjunctive, $t(47) = 9.27$, $p < .0001$ (exoneration).

Finally, Figure 4C shows that when reasoning about conjunctive causes, subjects also exonerated when predicting a cause in the absence of an effect: The cause was rated higher when the other cause was absent (32.1 vs. 19.8 when present). In contrast, the independent cause network exhibited a typicality effect in which a cause was rated as more likely when the other cause was present (26.9 vs. 18.8) even though those causes were conditionally inde-

pendent. Analyses of the data in Figure 4C revealed no main effects (both Fs < 1) but an interaction, $F(1, 47) = 15.43$, $MSE = 326$, $p < .001$. In particular, the presence of the other cause increased ratings for independent causes, $t(47) = 2.30$, $p < .05$ (typicality) but decreased them for conjunctive ones, $t(47) = 2.80$, $p < .01$ (exoneration).

Feature likelihood results. Finally, the feature likelihood ratings shown in Figure 4F exhibit the predicted interaction between network type and feature type shown in Figure 3F. Specifically, the conjunctive effect CE was rated to be less probable than both its causes (65.8 vs. 75.8) and the independent effect IE (76.8). The independent effect in turn was rated as more likely than its causes (70.9). A 2 × 2 ANOVA of these data revealed no main effects (Fs < 1) but an interaction, $F(1, 47) = 15.5$, $MSE = 108$, $p < .001$. The conjunctive effect was rated lower than its causes, $t(47) = 2.87$, $p < .01$, and the independent effect, $t(47) = 2.68$, $p = .01$; the independent effect in turn was rated higher than its causes, $t(47) = 2.56$, $p = .01$. Note that the latter result replicates previous research demonstrating a reverse causal status effect when an effect has multiple independent causes (Rehder, 2003a; Rehder & Hastie, 2001; Rehder & Kim, 2006; see Footnote 1).

Discussion

The first question asked by Experiment 1 was whether category membership judgments were sensitive to the functional relationship relating an effect feature to its causes. The answer is that they very much were. An object with an effect and one cause was rated to be a moderately probable category member when the causes were independent but not when causes were conjunctive; conversely, an object with an absent effect and one cause was rated a better category member when causes were conjunctive versus independent. The importance of coherence to classification involving an independent cause network replicates previous research (Rehder, 2003b; Rehder & Hastie, 2001); that a conjunctive cause network exhibits its own unique pattern of coherence (and one that is predicted by the model) represents a new finding.

A second question was whether the same representation of independent and conjunctive causes could account for performance on multiple tasks. The answer is that it can, as the predicted sensitivity to functional form also obtained for feature prediction and likelihood judgments. When one of two causes was present, inferences regarding the presence of an effect were very much higher for independent versus conjunctive causes. And, an effect was rated more probable when its causes were independent versus conjunctive causes.

As expected, one important prediction failure concerned the patterns of conditional independence predicted by the standard model. According to the standard model, the probability of a conjunctive effect should be the same regardless of whether zero or one cause is present (see Figure 3A), and the probability of an independent cause when the effect is absent should be independent of the state of the other cause (see Figure 3C), but subjects instead exhibited a typicality effect in which ratings increased as a function of the number of typical category features already present. The following section also demonstrates how the entire pattern of results in Experiment 1—including the typicality effect—can be reproduced by assuming that category features are related via additional causal relations.

Theoretical Modeling

To demonstrate to what extent the causal models in Figure 2A provide not only a qualitative but also a quantitative account of subjects' judgments, they were simultaneously fit to Experiment 1's inference, classification, and feature likelihood ratings. The c parameters associated with the four explicit causes in Figure 2A were assumed to be equal (i.e., $c_{IC1} = c_{IC2} = c_{CC1} = c_{CC2} = c$) as were all m parameters ($m_{IC1,IE} = m_{IC2,IE} = m_{CC1,CC2,IE} = m$), and both b parameters ($b_{IE} = b_{CE} = b$). In addition, a γ parameter was introduced for each task that applies a nonlinear power transformation of the probability derived from the CGM onto the rating scale. That transformed probability is then multiplied by 100 to bring it into the [0, 100] range of the ratings. That is, ratings were predicted as follows:

$$rating^{classification}(o_i) = 100p_k(o_i;c,\ m,\ b)^{\gamma^{classification}}$$

$$rating^{inference}(r_i \mid g_i) = 100p_k(r_i \mid g_i;c,\ m,\ b)^{\gamma^{inference}}$$

$$rating^{feature-likelihood}(f_i) = 100p_k(f_i;c,\ m,\ b)^{\gamma^{feature-likelihood}},$$

where g_i and r_i are the given and predicted features on inference trial i, respectively, o_i is the object categorized on classification trial i, and f_i is the feature presented on feature likelihood trial i. c, m, and b were free parameters constrained to the range [0, 1]. The γs were free parameters constrained to the range [0, 10].[4] This model was fit to each subject's 30 ratings by identifying parameters that minimized squared error.[5] The best fitting parameters averaged over subjects are presented in Table 3. In fact, this model was able to account for many of the qualitative effects seen in this experiment, including coherence effects in classification, explaining away and exoneration effects in inference, and the differences in the marginal probabilities of the independent and conjunctive effects. As a result, the model achieved a respectable correlation between its averaged predictions and the 30 empirical data points in Figure 4 (.955; .844 when correlations are computed for each subject and then averaged).

As expected, however, the standard model was unable to account for the typicality effect observed in the inference data. For example, it predicts that the probability of a conjunctive effect is the same regardless of whether zero or one cause is present, a result that is at odds with the right side of Figure 4A. And, it predicts that probability of an independent cause is the same when the effect is absent regardless of the state of the other cause, in contrast to subjects' judgments in the left side of Figure 4C.

Recall that Rehder and Burnett (2005) proposed that the typicality effect arises because people assume that category features are related

[4] Separate scaling parameters were required because of the very different range of probabilities associated with each task. As demonstrated by the example in Table 1, the classification (i.e., joint) probabilities are all relatively low, but subjects are likely to map those probabilities onto the rating scale in a manner that makes use of the upper half of the scale. For example, a value of .33 for $\gamma^{classification}$ maps probabilities in the range [0.01, 0.40] into the range [.22, .74]. Conversely, because each category feature was described as occurring in "most" category members, marginal probabilities are likely to be high, but subjects are likely to make use of the lower half of the scale. For example, a value of 3 for $\gamma^{feature-likelihood}$ maps the probability range [0.60, 0.99] into [.22, .97].

[5] Model fits were carried about by conducting an initial grid search of the parameter space followed by standard hill climbing using R's fminsearch procedure.

Table 3
Parameter Estimates From Experiments 1 and 2 for the Standard and Underlying Mechanism (UM) Models

Experiment 1			Experiment 2		
	Standard CGM	UM model		Standard CGM	UM model
Parameter			Parameter		
c	.736 (.027)	.505 (.038)	c	.802 (.014)	.599 (.025)
m	.819 (.026)	.704 (.042)	m_w	.654 (.022)	.447 (.034)
			m_s	.816 (.018)	.677 (.031)
b	.337 (.046)	.148 (.033)	b	.462 (.029)	.270 (.030)
c_{UM}		.609 (.050)	c_{UM}		.680 (.035)
m_{UM}		.763 (.052)	m_{UM}		.729 (.034)
$\gamma^{classification}$	0.29 (.015)	0.26 (.016)	$\gamma^{classification}$	0.29 (.020)	0.26 (.011)
$\gamma^{inference}$	3.73 (.528)	3.82 (.523)	$\gamma^{inference}$	3.99 (.318)	4.29 (.330)
$\gamma^{feature\text{-}likelihood}$	2.17 (.368)	4.11 (.587)	$\gamma^{feature\text{-}likelihood}$	2.50 (.294)	3.84 (.367)
Measures of fit					
Avg. R	.844	.895		.854	.900
Avg. *AIC*	182.1	174.1		176.9	170.3

Note. CGM = causal graphical model; AIC = Akaike's information criterion. Standard errors are in parentheses.

via a common UM. To test this account, the UM model in Figure 2B was fit to the results of Experiment 1 by introducing two additional free parameters. Recall that c_{UM} is the probability that UM is present, and m_{UM} is the power of the causal links between it and the six category features. The best fitting parameters for the UM model are presented in Table 3 and its predictions are shown in Figure 4, superimposed on the empirical data. The figure shows that this model not only accounted for the same effects as the standard model (coherence, explaining away, etc.), it also accounted for the typicality effect; that is, it reproduced subjects' larger inference ratings when predicting an effect given one versus zero causes in the conjunctive condition (see the right side of Figure 4A) and when predicting a cause when the effect was absent given one versus zero causes in the independent condition (see the left side of Figure 4C).

Figure 4 also reveals some differences between the predicted and observed ratings. In the classification results, it underpredicts subjects' ratings on those trials in which one cause and one effect is present (see Figure 4D) and overpredicts ratings in which causes and effects are all absent (see Figure 4E). It also overpredicts the likelihood of the independent effect (see Figure 4F). Model fitting and individual-difference analyses presented as part of Experiment 2 considers potential reasons for these mispredictions. Nevertheless, the UM model's ability to account for the qualitative effects in Figure 4 resulted in a correlation between the observed and predicted ratings of .984 (.895 averaged over subjects). Note that the better fit of the UM model relative to the basic model was also reflected in a measure (Akaike's information criterion [*AIC*]) that takes into account its extra two parameters (174.1 averaged over subjects vs. 182.1).[6]

A number of variants of the UM model were explored. First, to assess the claim that the differences observed between independent and conjunctive causes can be accounted for solely by changing the functional relationship between causes and effect; a more general version of the model in which each network had its own set of causal model parameters (i.e., its own c, m, and b) was fit to the data. This model achieved a higher correlation with the observed data relative to the basic UM model, of course (.987 vs. .984), but its higher *AIC* (176.2 vs. 174.1) revealed its poorer fit, taking into account its three extra parameters. Second, a version of the UM model was defined that used a linear rather than a noisy-or integration rule.[7] However, its worse fit to the data (*AIC* = 178.1) indicates that the noisy-or formulation provides a better characterization of the category-based inferences. Finally, to assess the claim that the same causal model can account for performance on three different tasks (prediction, classification, and feature likelihood judgments), I fit a version of the model generalized to have separate causal model parameters (i.e., a separate c, m, b, c_{UM}, and m_{UM}) for each task. This model achieved a higher correlation (.992), but its higher *AIC* (180.6) revealed its poorer fit, taking into account its extra parameters. Nevertheless, analyses presented in Experiment 2 reveal some potential differences regarding how causal knowledge affects category membership decisions as compared with the other two types of judgments.

[6] $AIC = n\log(SSE/n) + 2(p + 1)$ where SSE = sum of squared error for a participant, n = number of data points fit (30), and p = a model's number of parameters (6 and 8 for the models in Figures 2A and 2B, respectively). This measure was deemed by Burnham and Anderson (1998) as appropriate for comparing models fit by least squares.

[7] A well-known alternative to the noisy-or integration rule is a linear rule in which multiple causal influences are summed. For example, the integration function for the independent and conjunctive cause networks represented by Equations 4 and 5 can be replaced with,

$$p_k(IE = 1|IC_1, IC_2) = 1/(1 + exp(-(ind(IC_1)w_{IC_1,IE} + ind(IC_2)w_{IC_2,IE} + w_{IE})))$$

$$p_k(CE = 1|CC_1, CC_2) = 1/(1 + exp(-(ind(CC_1)ind \times (CC_2)w_{CC_1,CC_2,CE} + w_{CE}))),$$

where $w_{IC_1,IE}$ and $w_{IC_2,IE}$ represent the strength of the causal relations between IE and IC_1 and IC_2, respectively, $w_{CC_1,CC_2,CE}$ represents the strength of the conjunctive relationship, and w_{IE} and w_{CE} represents the strength of alternative causes of IE and CE. These equations can be further generalized to represent the UM model in Figure 2B and the Appendix, and the result can then be used to derive joint distributions for the independent and conjunctive cause networks in the same manner as for the noisy-or integration rule. Full details of the fits of this model are available from the author.

Experiment 2

The aim of Experiment 2 was to conduct a second assessment of reasoners' sensitivity to independent versus conjunctive causes and provide additional tests of the standard and UM models. In addition to the differences in the patterns of inferences shown in Figure 3, the model predicts that inferences based on independent and conjunctive cause networks will respond very differently to changes in the *strengths* of the causal relationships. Experiment 2 tests these predictions by manipulating causal strength as a within-subjects variable and network type as a between-subjects variable. All subjects learned one of the six-feature categories used in Experiment 1. Half the subjects learned that those six features formed two independent cause networks, as shown in Figure 6A. The other half learned that they formed two conjunctive cause networks (see Figure 6B). For each subject, the causal relation(s) of one subnetwork were described as weak by saying that they "occasionally" produced the effect. The relation(s) of the other subnetwork were described as strong by saying that they "often" produced the effect.

The predictions made by the networks in Figure 6 are presented in Figure 7. Values of .33 and .67 were assumed for the causal power *m* parameters for the weak and strong subnetworks, respectively. The remaining parameters are set to the same value as in Figure 2A (*c*s = .67 and the *b*s = .20). Figures 7D and 7E reveal that the patterns of coherence exhibited by each type of network interact in distinct ways with causal strength. In Figure 7D in which the effect is present, the independent cause network exhibits stronger coherence effects (category membership is more likely) as a function of causal strength when either one or two causes are present, whereas the conjunctive cause network does so only when both causes are present. Conjunctive causes exhibit this pattern because the causal mechanism is inoperative when only one cause is present; thus, increasing causal strength in this case does not contribute to coherence. Figure 7D thus implies a three-way interaction between network type, strength, and the number of causes. Analogously, in Figure 7E (in which the effect is absent), the independent cause network exhibits stronger coherence effects (category membership is less likely) as a function of causal strength when either one or two causes is present, whereas the conjunctive cause network does so only when both causes are present.

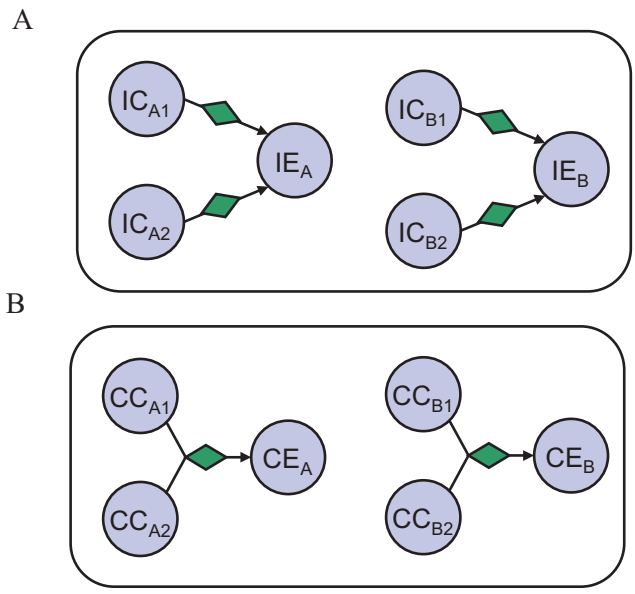

Figure 6. A schematic representation of the category features and causal relations learned in Experiment 2. IC = independent cause; IE = independent effect; CC = conjunctive cause; CE = conjunctive effect. See the online article for a color version of this figure.

Figures 7A–C present how conditional probability judgments are affected by causal strength. First, Figure 7A shows the unsurprising result that both the probability of an effect given both causes is higher for stronger versus weaker causal relations. However, when only one cause is present, the probability of the effect increases with causal link strength for independent causes only. Next, Figure 7B shows that whereas the (signed) difference in the probability of a cause given the effect when the other cause is present versus absent *decreases* as causal strengths increase when causes are independent (i.e., explaining away gets stronger), it instead *increases* when causes are conjunctive (exoneration gets stronger). In Figure 7C, only the conjunctive cause network exhibits an interaction between strength and the state of the other cause. Just like Figures 7D and 7E, Figures 7A–C thus each imply a three-way interaction involving network type and causal strength.

Finally, Figure 7F reveals that both the independent and the conjunctive effect should become more probable as the strengths of its causes increases. They should do so because they are more likely to be generated by a stronger causal mechanism.

Method

Materials. Subjects learned one of the six categories used in the first experiment. For subjects in the independent cause condition, the causal links of one subnetwork were described as weak and those of the second subnetwork were strong. For example, to convey a weak causal relation between a damaged fan belt and a long-lived generator in Rogos, they were told "A damaged fan belt occasionally causes a long-lived generator." To convey a strong relation, *often* was used instead of *occasionally*. Similarly, subjects in the conjunctive cause condition learned one weak and one strong relation (e.g., "Butane-laden fuel and loose fuel filter gaskets occasionally/often cause a hot engine temperature").

The assignment of the six typical category features to subnetworks was balanced over subjects such that for each category, one triple of features formed the weak subnetwork and the other formed the strong subnetwork for half the subjects, and this assignment was reversed for the other half.

Procedure. The procedure was identical to Experiment 1 except for the information about the strengths of the causal links and questions on the multiple-choice test that queried subjects about those strengths.

Participants. Ninty-six New York University undergraduates received course credit for participating in this experiment. There were four between-subject factors: whether subjects learned independent or conjunctive causes, the assignment of physical features to the weak or strong subnetworks, the two task presentation orders (whether prediction or classification was first), and which of the six categories was learned. Participants were randomly assigned to

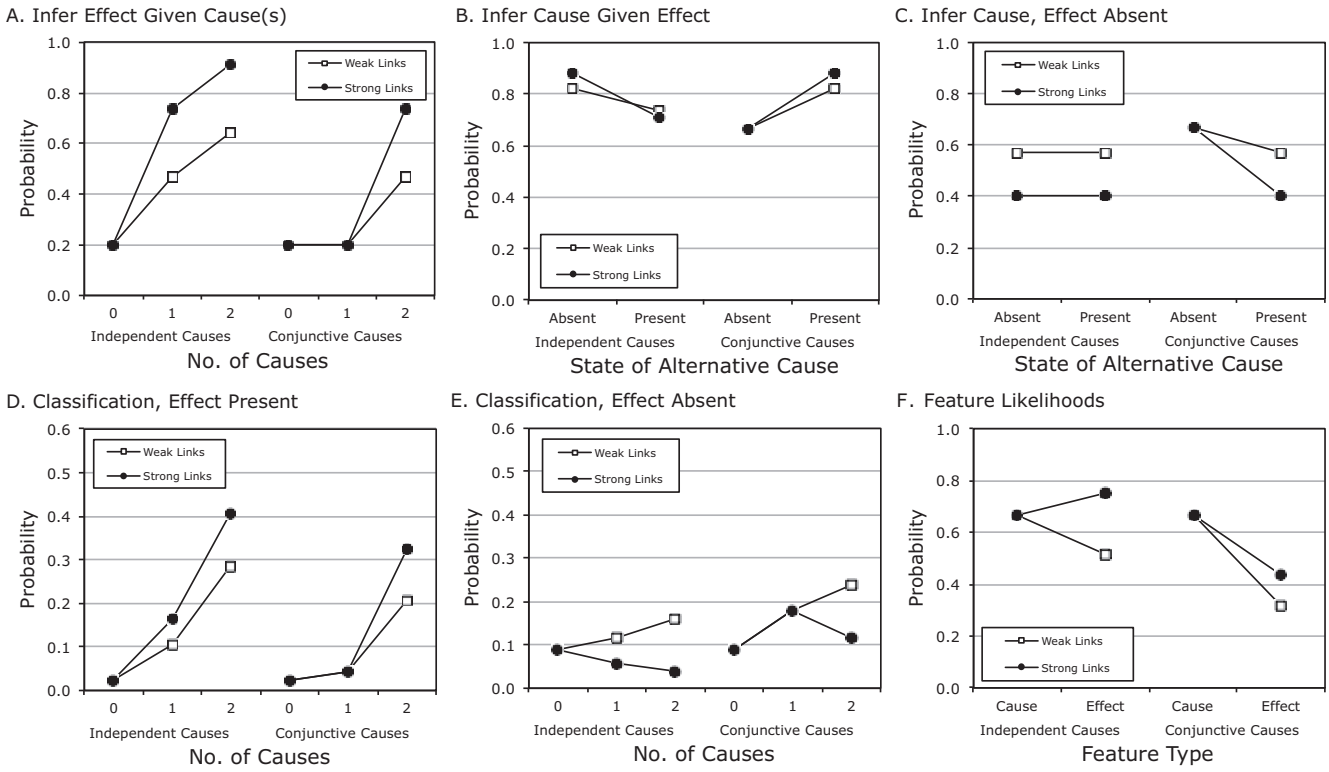

Figure 7. Predictions for the standard independent and conjunctive cause networks of Experiment 2 as a function of whether the causal relations are weak or strong. Inference predictions are in Panels A–C, classification predictions are in Panels D–E, and feature likelihood predictions are in Panel F.

these $2 \times 2 \times 2 \times 3 = 24$ between-participant cells subject to the constraint that an equal number appeared in each cell.

Results

Initial analyses revealed no effects of which category subjects learned, the assignment of features to causal roles (i.e., the weak or strong subnetwork), or task presentation order, and so the results are presented collapsed over these factors.

Classification results. Classification ratings are presented in Figures 8D and 8E. First, when the effect was present (see Figure 8D), classification judgments revealed an increase in coherence effects when causal links were stronger, albeit a modest one (e.g., ratings of 93.7 vs. 87.6 averaged across both networks when two causes were present). Moreover, for the conjunctive cause network, this effect obtained only when both causes were present, as predicted in Figure 7D. A $3 \times 2 \times 2$ ANOVA of these data was conducted. As in Experiment 1, main effects of the number of causes, $F(2, 188) = 411.48$, $MSE = 436$, $p < .0001$; network type, $F(1, 94) = 5.51$, $MSE = 273$, $p = .02$; and an interaction between the two, $F(2, 188) = 6.92$, $MSE = 436$, $p < .01$, confirmed the distinct patterns of coherence exhibited by the two types of networks. In addition, the predicted main effect of causal strength, $F(1, 94) = 7.91$, $MSE = 155$, $p < .01$, and three-way interaction between number of causes, network, and strength, $F(2, 188) = 4.95$, $MSE = 134$, $p < .01$, obtained. In a separate 3×2 ANOVA of the conjunctive cause network, strength interacted with the number of causes in the manner predicted, as seen on the right-hand side of Figure 7D, $F(2, 94) = 6.85$, $MSE = 174$, $p < .01$. The corresponding interaction for the independent cause network shown on the left-hand side of Figure 7D did not obtain in the empirical data, however ($F < 1$).

In comparison to Figure 8D, Figure 8E shows a more prominent effect of the causal strength manipulation, especially for the independent cause network. A $3 \times 2 \times 2$ ANOVA revealed a main effect of the number of causes, $F(2, 188) = 10.28$, $MSE = 925$, $p < .001$. Although the two-way interaction between network type and the number of causes was marginal, $F(2, 188) = 2.22$, $MSE = 925$, $p = .12$; the more focused test of the interaction between network type and the quadratic trend of the number of causes reached significance, $F(1, 94) = 4.57$, $MSE = 677$, $p = .03$, corroborating the distinct patterns of coherence supported by independent versus conjunctive causes first observed in Experiment 1. The analysis also revealed a main effect of strength, $F(1, 94) = 10.63$, $MSE = 215$, $p < .01$, and strength interacted with network type, $F(1, 94) = 6.57$, $MSE = 215$, $p = .01$, and the number of causes, $F(1, 94) = 12.08$, $MSE = 192$, $p < .0001$. Separate 3×2 ANOVAs of the two network types confirmed that each was sensitive to causal strength (both ps $< .04$). The three-way interaction did not reach significance, $F(2, 188) = 1.54$, $MSE = 192$, $p = .22$.

Feature inference results. Feature inference ratings are presented in Figures 8A–C. In general, these ratings reflect the qual-

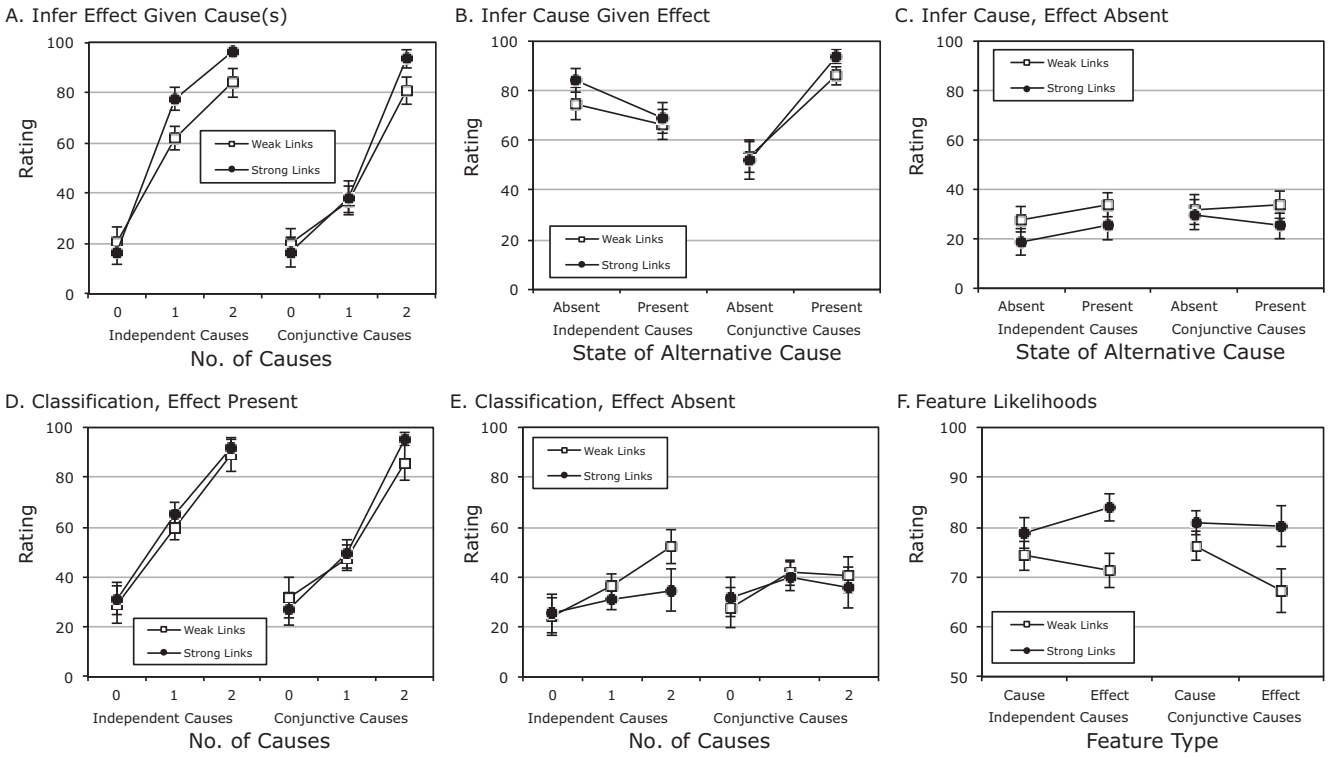

Figure 8. Empirical results from Experiment 2. Inference results are in Panels A–C, classification results are in Panels D–E, and feature likelihood results are in Panel F. Error bars are 95% confidence intervals.

itative predictions shown in Figures 7A–C, albeit with the one misprediction seen in the first experiment. Starting with Figure 7A, the pattern of forward predictions again reflected subjects' sensitivity to the functional relationship between causes and effect, as the effect was rated more likely to present when one cause was present for the independent (average rating of 69.7) as compared with conjunctive (37.8) causes. More interestingly, when both causes were present, inference ratings were larger for strong causal links (94.8) versus weak ones (82.2), confirming that subjects were sensitive to the manipulation of causal strength. In contrast, ratings were unaffected by causal strength when both causes were absent (16.7 vs. 20.4). Most importantly, when only one cause was present, inferences were sensitive to causal strength for the independent (77.7 vs. 61.8) but not the conjunctive network (38.4 vs. 37.1), consistent with the predictions in Figure 7A.

A $3 \times 2 \times 2$ ANOVA of the data in Figure 7A revealed main effects of the number of causes, $F(2, 188) = 501.72$, $MSE = 469$, $p < .0001$; network type, $F(1, 94) = 38.56$, $MSE = 515$, $p < .0001$; and causal strength, $F(1, 94) = 16.68$, $MSE = 295$, $p < .0001$. The number of causes interacted with network type, $F(2, 188) = 31.53$, $MSE = 469$, $p < .0001$, as in Experiment 1, and, in a new result, also with causal strength, $F(2, 188) = 22.89$, $MSE = 151$, $p < .0001$. Finally, this analysis yielded the predicted three-way interaction between number of causes, network type, and causal strength, $F(2, 188) = 5.94$, $MSE = 151$, $p < .01$. As in Experiment 1, ratings in the conjunctive cause condition differed from the theoretical predictions in that the effect was rated as more probable in the presence of one cause (37.8) versus none (18.4), $t(47) = 7.07$, $p < .0001$. That is, subjects again exhibited a typicality effect.

Regarding ratings of the conditional probability of a cause, Figure 8B reveals the same interaction between the state of the other cause and network type seen in Experiment 1, namely, subjects exhibited explaining away for independent causes (ratings of 79.7 when the other cause was absent vs. 68.0 when present) and exoneration for conjunctive causes (52.9 vs. 90.0). More importantly, both of these effects were stronger for stronger causal relations, consistent with the predictions in Figure 7B. A $2 \times 2 \times 2$ ANOVA of these data revealed an effect of the state of the other cause, $F(1, 94) = 20.55$, $MSE = 767$, $p < .0001$; no effect of network type ($F < 1$); but an interaction between these two factors, $F(1, 94) = 75.03$, $MSE = 767$, $p < .0001$. The new finding is the fact that this two-way interaction itself interacted with causal strength, $F(1, 94) = 18.03$, $MSE = 88$, $p < .0001$.

Turning to Figure 8C, when the effect was absent, independent cause subjects were less likely to infer the presence of a cause for stronger versus weaker causal links, an effect that was anticipated by Figure 7C. Those inferences also exhibited a small typicality effect in which inferences were sensitive to the state of the alternative cause. For the conjunctive cause condition, one surprising result is that subjects showed a markedly weaker exoneration effect than in Experiment 1. Model fitting results presented below suggest that this result stems from the weaker causal links used in this experiment as compared with the first. A $2 \times 2 \times 2$ ANOVA of these data revealed an effect of neither the state of the other cause, $F(1, 94) = 1.97$, $MSE = 338$, $p = .16$, nor network type,

$F(1, 94) = 1.53$, $MSE = 906$, $p = .22$, but an interaction between these two factors, $F(1, 94) = 4.39$, $MSE = 338$, $p = .04$, replicating Experiment 1. In addition, there was a main effect of causal strength in the predicted direction, $F(1, 94) = 24.14$, $MSE = 185$, $p < .0001$. The predicted three-way interaction between strength and the other two factors was only marginal, however, $F(1, 94) = 2.27$, $MSE = 133$, $p = .14$.

Feature likelihood results. The feature likelihood ratings shown in Figure 8F exhibit the same interaction between network type and feature type found in Experiment 1 in which the conjunctive effects were rated as less prevalent than the independent effects. In addition, the figure confirms the prediction (see Figure 7F) that effects become more prevalent as causal links get stronger. A 2 × 2 × 2 ANOVA of these data replicated Experiment 1's interaction between causal role (cause or effect) and network type, $F(1, 94) = 5.65$, $MSE = 147$, $p = .02$. The new finding is that causal role also interacted with causal strength, $F(1, 94) = 37.90$, $MSE = 42$, $p < .0001$. Separate 2 × 2 ANOVAs revealed that the effect was rated as more prevalent relative to the causes when causal links were strong versus weak for both the independent, $F(1, 47) = 19.17$, $MSE = 41$, $p < .0001$, and conjunctive networks, $F(1, 47) = 18.75$, $MSE = 43$, $p < .0001$. The former result replicates previous research showing that the difference in the importance of causes and effects (i.e., the causal status effect) grows smaller as the strength of the causal links increase (see Foonote 1 and Rehder & Kim, 2010), whereas the latter extends this finding to conjunctive causes. A three-way interaction was neither predicted for these data nor found ($F < 1$).

Theoretical Modeling

Following Experiment 1, the causal models on which subjects were instructed were fit to the empirical ratings. Independent cause networks were fit for subjects taught the causal structures in Figure 6A, whereas conjunctive cause networks were fit for those taught the structures in Figure 6B. To capture the effects of the within-subject causal strength manipulation, two strength parameters, m_w and m_s, were used to model inferences for the weak and strong networks, respectively. This model, whose best fitting parameters are presented in Table 3, was not only able to account for the same effects it did in Experiment 1 (coherence, explaining away, exoneration, etc.), it also reproduced many of the qualitative effects of causal strength, achieving a correlation of .958 (.854 averaged over subjects) between its predictions and the 60 empirical data points in Figure 8. But because it was unable to reproduce a typicality effect, a version of the models in Figure 6 with a common UM analogous to that in Figure 2B was also fit. The parameter estimates for this model (see Table 3) are similar to those in Experiment 1, with the exception of the weaker causal strength parameters (.447 and .680 in the weak and strong conditions vs. .704 in the previous experiment), apparently reflecting the use of *occasionally* and *often* to describe those relationships in Experiment 2. The fits of this model are shown in Figures 9 and 10 superimposed on the empirical results for the independent and conjunctive cause conditions, respectively. The left side of Figure 9C and the right side of 10A reveal that this model was able to reproduce the

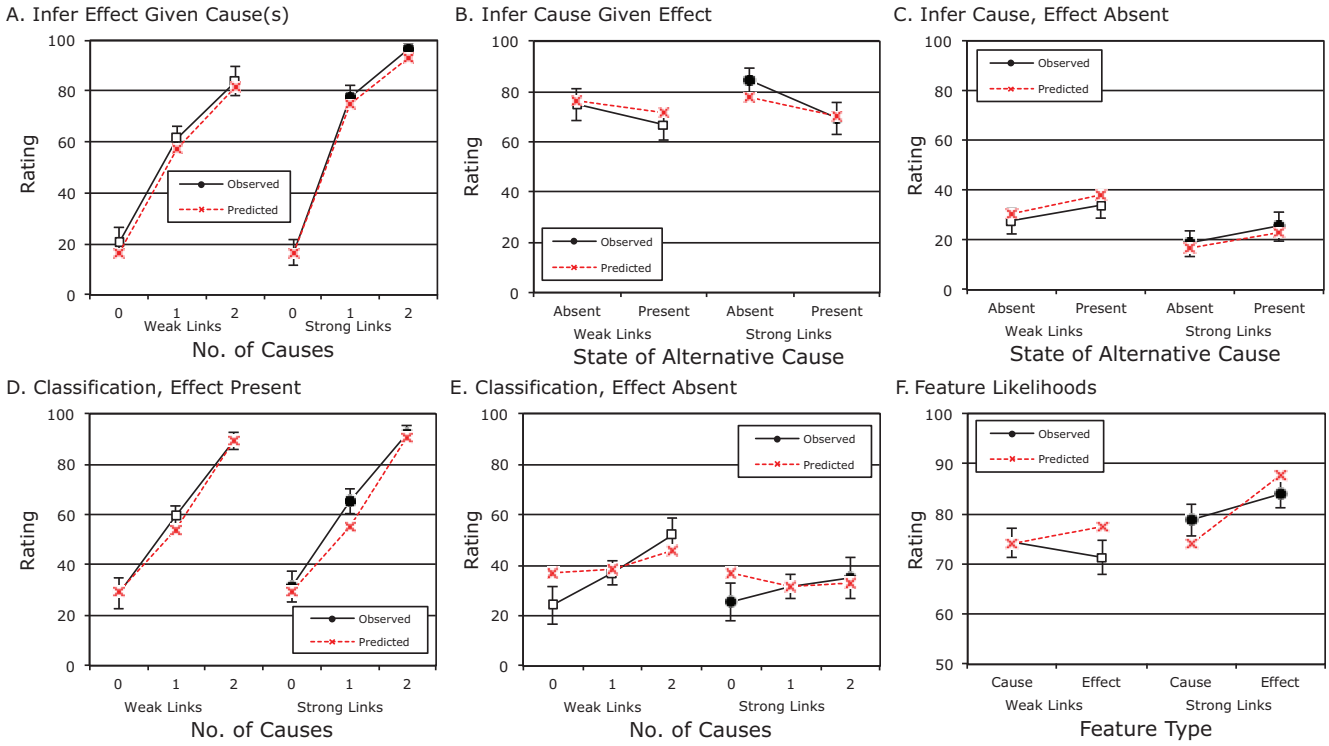

Figure 9. Fits of the underlying mechanism model superimposed on the empirical data of the independent cause condition of Experiment 2. Error bars are 95% confidence intervals. See the online article for a color version of this figure.

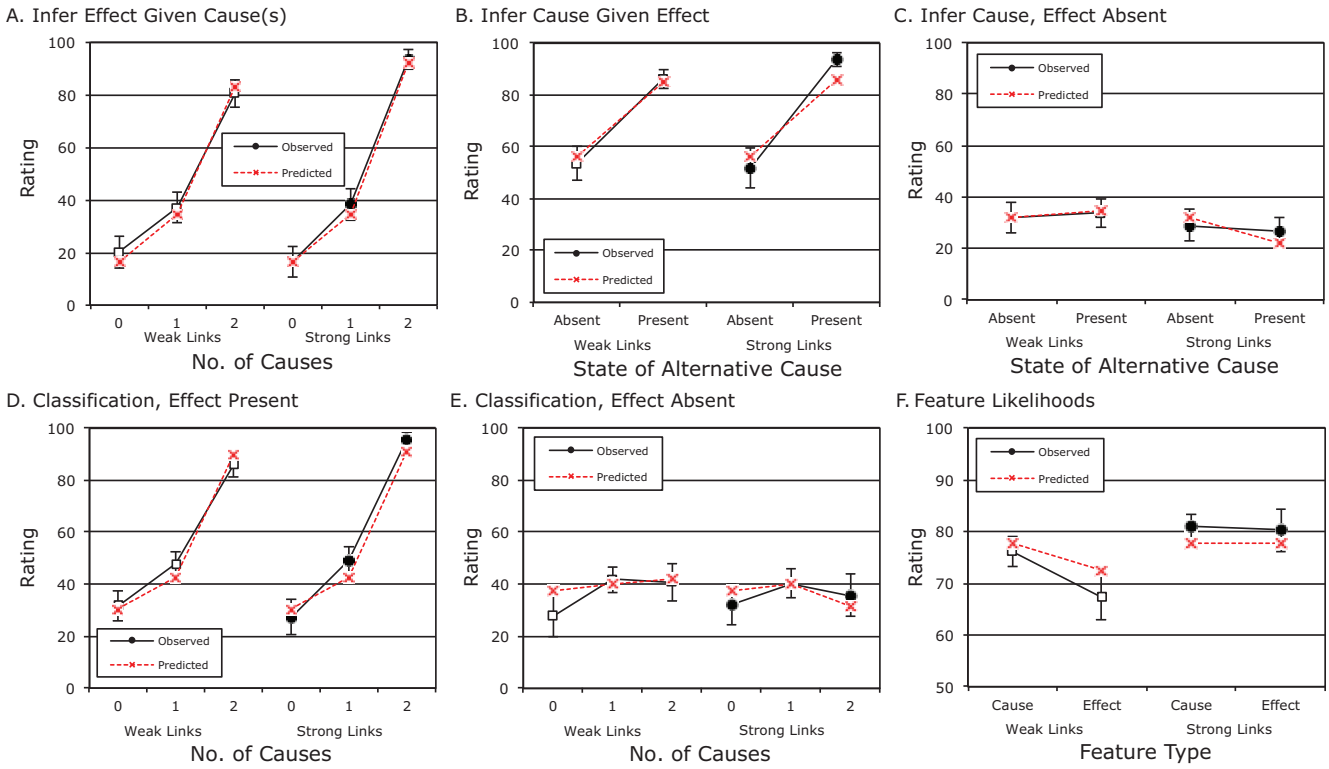

Figure 10. Fits of the underlying mechanism model superimposed on the empirical data of the conjunctive cause condition of Experiment 2. Error bars are 95% confidence intervals. See the online article for a color version of this figure.

typicality effect, just as it did in Experiment 1. The model's ability to account for most of the trends in Figure 8 resulted in a correlation between the observed and predicted ratings of .985 (.900 averaged over subjects). The better fit of the UM model relative to the basic model was also reflected in *AIC*s that account for its two additional parameters (170.3 averaged over subjects vs. 176.9).

As in Experiment 1, a number of generalizations of the UM model were tested. First, to assess the claim that the differences observed between the weak and strong subnetworks were accounted for solely by changing the causal power parameter (m_w vs. m_s), a version of the model in which each subnetwork had its own set of parameters (i.e., its own c, b, and m_{UM} in addition to its own m) was fit to the data. This model achieved a poorer fit than the basic UM model, taking into account its three extra parameters (*AIC*s of 173.1 vs. 170.3). Second, to assess whether subjects' inferences were better described by a linear integration rule, a linear version of the UM model was fit (see Footnote 7), but, as in Experiment 1, it yielded a worse fit relative to the UM model ($AIC = 171.3$ vs. 170.3). Third, to assess the claim that the same causal model can account for performance on three different tasks (classification, prediction, and feature likelihood judgments), it was generalized to have separate causal model parameters (i.e., a separate c, m_w, m_s, b, c_{UM}, and m_{UM}) for each task. Again, this model yielded a poorer fit, taking into account its extra parameters ($AIC = 182.0$). Yet, some qualitative differences between tasks existed, as now discussed.

Assessing Individual Differences

Figures 9 and 10 reveal that the UM model yielded some of the same mispredictions seen in Experiment 1 (e.g., the overprediction of classification ratings when causes and effects are all absent). To assess whether subgroups of subjects who responded qualitatively different from one another contributed to these mispredictions, a single *k-means* clustering procedure was run on the fitted parameter estimates for the 144 subjects tested in this study (48 from Experiment 1 and 96 from Experiment 2). To equate the number of parameters, the weak causal strength parameter for Experiment 2 subjects (m_w) was omitted. Evidence for two clusters was strong and that for three intermediate. The following analysis is based on the three-cluster solution, consisting of 88 (Cluster 1), 38 (Cluster 2), and 18 (Cluster 3) subjects. Clusters 1 and 2 were most sharply distinguished on the parameters associated with the UM: Whereas the average values for parameters c_{UM} and m_{UM} were .58 and .70 in Cluster 1, respectively, they were .86 and .87 in Cluster 2. In addition, these clusters differed in their overall quality of fit: The mean absolute deviation between the observed and predicted values was 2.5 in Cluster 1 as compared with 5.4 in Cluster 2. Because the mispredictions in Cluster 2 largely occurred on the classification trials, the observed and predicted values for those trials are presented in Figure 11 for both clusters. (In calculating their contribution to Figure 11, the inferences of each Experiment 2 subject were averaged over the weak and strong conditions.) Differences between the classification ratings are especially prom-

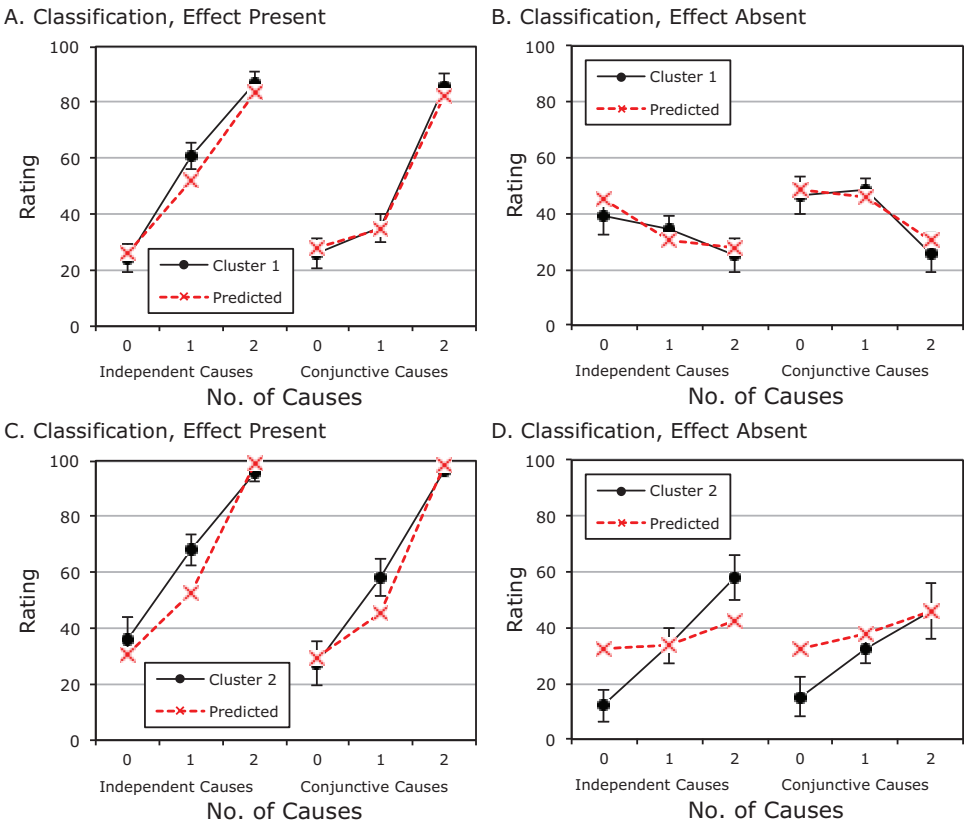

Figure 11. The classification ratings for two clusters of subjects from the 144 subjects tested in Experiments 1 and 2. A: For Cluster 1 ($N = 88$) when the effect was present. B: For Cluster 1 when the effect was absent. C: For Cluster 2 ($N = 38$) when the effect was present. D: For Cluster 2 when the effect was absent. The averaged predictions of the underlying mechanism model are presented superimposed on the empirical data points. Error bars are 95% confidence intervals. See the online article for a color version of this figure.

inent when the effect was absent: Whereas Cluster 1 exhibited a strong coherence effect (ratings decrease as more causes are present; see Figure 11B), Cluster 2 exhibited the reverse pattern (see Figure 11D). That is, Cluster 2's classification judgments were strongly influenced by typicality and only weakly so by causal structure (including the independent and conjunctive functional forms). Importantly, though, the other (feature prediction and likelihoods) judgments of these subjects showed sensitivity to causal structure and functional form, that is, were less dominated by typicality. As a result, the UM model yields a relatively poor fit for these individuals because it cannot account for the varying influence of typicality across the three tasks. Note that Cluster 2 accounts for the mispredictions cited earlier, namely, the underprediction of classification ratings when one cause and one effect is present (compare Figure 11C with Figures 4D, 9D, and 10D) and their overprediction when causes and effects are all absent (compare Figure 11D with Figures 4E, 9E, and 10E). The responses of the minority Cluster 3 fell roughly in between those of Clusters 1 and 2.

Discussion

The purpose of Experiment 2 was to assess whether the changes in independent and conjunctive inferences as a function of causal strength would obtain. The answer is that they (almost) all did. The effect of causal strength not only went in the predicted direction; in every case, it displayed the predicted three-way interaction involving the type of network (albeit that interaction was marginal in Figure 8C and nonsignificant in 8E). The pair of two-way interactions predicted for feature likelihood judgments (see Figure 7F) also obtained. In addition, theoretical modeling showed that a CGM representation of causal relations codifying those functional relationships but also enhanced with additional hidden causal relations provided a respectable quantitative fit. An analysis of individual differences revealed that, for some subjects, classification judgments are more vulnerable to typicality effects than the other two sorts of judgments.

General Discussion

The Role of Functional Form in Category-Based Judgments

The first question asked in this research was whether category membership judgments are sensitive to the different functional relationships that can relate an effect to its causes. The answer is that they very much are. For example, in Experiment 1, an object

with an effect and one out of two independent causes were assigned an average classification rating of 65 (on a 0–100 scale), indicating that it was considered to be a moderately probable category member. But the same item was given a rating that was more than 25 points lower when the causes were conjunctive. The importance of coherence to classification involving an independent cause network replicates previous research (Rehder, 2003b; Rehder & Hastie, 2001); the fact that a conjunctive cause network exhibits its own unique pattern of coherence (and one predicted by the proposed model) represents a new finding.

The causal models that accounted for category membership ratings also accounted for two other types of judgments (with the one exception discussed in the following section). For example, when causes were independent, subjects exhibited explaining away, that is, judged that a cause feature was less likely to be present when another cause was present, consistent with the numerous demonstrations of explaining away in the social psychology (e.g., Jones & Harris, 1967; McClure, 1998; Morris & Larrick, 1995) and cognitive (e.g., Khemlani & Oppenheimer, 2011; Oppenheimer, 2004; Rehder & Burnett, 2005) literatures. In contrast, when causes were conjunctive, subjects exhibited exoneration effects. This is the first demonstration that exoneration effects are entailed by a generative representation of conjunctions and that human reasoners exhibit that effect. The model also successfully predicted that an effect feature would be rated as less prevalent when it was the product of two conjunctive versus independent causes. These results demonstrate how a causal model can be used to simultaneously account for multiple types of category-based judgments.

This finding is important because it is well known that explanations of behavior that appeal to complex cognitive representations must face the problem of identifiability, the possibility that that behavior arises from other sources, such as the mental processes invoked by the task (Anderson, 1990). On one hand, many so-called *rational models* of cognition avoid this issue by claiming to provide only a computational level account (in Marr's familiar terms), that is, one that specifies the cognitive function that is being computed without making claims regarding mental processes or representations. However, because a defining property of any mental representation is that it is accessible to multiple mental processes, evidence for the psychological reality of a representation can accrue through converging operations, that is, by showing that the same representation accounts for performance on multiple tasks. In this light, the present research provides evidence for the more ambitious claim that the proposed generative representations of independent and conjunctive causes correspond to the actual mental representations that reasoners use to render category-based judgments (cf. Jones & Love, 2011).

Experiment 2 provided an especially stringent test of the causal model framework by providing subjects with information corresponding to the strength of the causal relations. In fact, each of the aforementioned effects varied with causal strength in the predicted manner: As causal strength decreased, coherence effects in classification became weaker, explaining away and exoneration in feature inference became less prominent, and effect features were rated as less prevalent. Moreover, quantitative model fits yielded generally good accounts of the observed ratings. Whereas parametric variations of a category's causal model and quantitative model fits have been conducted for other network topologies (a three-element chain network in Rehder & Kim, 2010), these experiments are the first to do the same with common effect networks (with either independent or conjunctive causes). Nevertheless, it is important to note that an individual-differences analysis suggested that classification judgments might be more prone to exhibit the typicality effects discussed below than the other two sorts of judgments.

As mentioned, a conjunctive relation is just one of the ways that multiple causes can interact to produce their effect. A disabler may, when present, disrupt the normal operation of a cause–effect relation (Park & Sloman, 2013; Walsh & Sloman, 2004, 2008). In contrast, an inhibitor might prevent any occurrence of an effect. And, a causal relation might be preventative rather than generative (making its effect less likely). Each of these functional forms implies a distinct joint distribution and so distinct patterns of classification, feature inference, and feature likelihood judgments. Each thus represents an important avenue for future research.

In summary, Experiments 1 and 2 provide further evidence in favor of the causal model view of conceptual representation. This approach has received support from yet other sorts of category-based judgments. Oppenheimer, Tenenbaum, and Krynski (2013) and Rehder and Kim (2009) have assessed the use of causal models when classification amounts to a case of diagnostic reasoning, that is, when one reasons from evidence (e.g., medical symptoms) to their underlying cause (a disease). Rehder (2009) showed how the probabilistic noisy-or representation of causal relations can account for people's category-based inductions—generalizing a new feature to a whole category—when that feature is causally related to existing ones (see also Rehder, 2006). In the domain of category learning, Waldmann, Holyoak, and Fratianne (1995) found that the match between interfeature correlations in learning data and those expected on the basis of a category's causal model affected learning speed and the category representation that was ultimately acquired. And, Kim and Keil (2003) and Martin and Rehder (2013) have investigated how causal models predict what information reasoners should choose in order to make an accurate classification. But despite its success, there was also a prominent exception to the a priori predictions of the causal model approach, as now discussed.

Hidden Mechanisms in Category Representations

It is important to emphasize again that subjects' inferences were not in full accord with predictions derived from the causal models on which they were instructed—in particular, subjects' feature inferences exhibited a typicality effect in which inferences were stronger whenever more features were present, regardless of whether those features were (according to the Markov condition) independent of the feature being predicted. Whereas the typicality effects seen with the current independent cause network replicates Rehder and Burnett (2005); those observed with the conjunctive cause network extend the generality of the typicality effect to a new type of network.

These findings corroborate Rehder and Burnett's (2005) proposal regarding people's assumption about categories' UM. As mentioned, the assumption of a UM can be viewed as representing the beliefs that people have about the hidden causal processes associated with categories (Gelman, 2003; Medin & Ortony, 1989). On this account, an object that already has many or most of

the category's characteristic features will be considered a "well functioning" category member, which in turn will lead a reasoner to assume it also has characteristic values on its unobserved feature dimensions. Conversely, a category member with several uncharacteristic features will be viewed as having mechanisms that are operating in ways that are unexpected for members of that kind, leading one to suspect that it will be unusual on other dimensions as well. In fact, the causal models on which subjects were instructed were able to reproduce the typicality effect when they were enhanced with a representation of a UM that links category features.

Violations of independence have been observed in causal reasoning studies not involving category features, and so one can ask whether the explanations offered for those violations are applicable here. First, Walsh and Sloman (2008; see also Park & Sloman, 2013; Walsh & Sloman, 2004) asked subjects to reason about real-world vignettes involving three variables related by causal knowledge into a common cause network. They found that subjects' inferences reflected nonindependence among the effect features even when the state of the common cause was known, violating the Markov condition. Walsh and Sloman attributed this result to reasoners inducing the presence of shared disabling conditions between causal links, allowing one to infer the absence of one effect given the absence of another (even when the common cause was present) on the grounds that if one causal link was disabled, it was likely the other one was too. Second, Mayrhofer et al. (2010; see also Mayrhofer & Waldmann, 2013) instructed subjects on scenarios involving mind-reading aliens connected in a common cause network and found that the thought of the "effect" aliens were not independent conditioned on the common cause. However, the size of the violation varied with instructions that manipulated whether the "common cause" alien was the "sender" or "receiver" of the thoughts, with violations much weaker in the latter condition, a result the authors interpreted as reflecting the influence of perceived agency on the causal model that subjects reasoned with. In the sending condition, it is natural to assume that the sender's ability to transmit thoughts to the effect aliens relied on some sort of common mechanism, an assumption that again undermines the expectation of conditional independence among the effects.

However, while accounting for independence violations with common cause networks, the sort of elaborations of causal relations proposed by Walsh and Sloman (2004; shared disablers) and Mayrhofer et al. (2010; a shared mechanism) provide no account of the violations seen here with causal networks involving two causes and a common effect. For example, they do not explain why the present subjects treated two independent causes as conditionally dependent when their effect was absent.

A more viable alternative account of the typicality effect comes from Rehder (2014), who tested how people reason with a number of causal network topologies involving variables drawn from various content domains (e.g., economics, meteorology, etc.). In fact, inferences with an independent cause network exhibited the same violations found in the present experiments. But an "underlying mechanism" account of those violations was less plausible given that the variables were not category features (and the counterbalancing of which variable states were described as causally related). I also found that the inferences of a substantial minority of subjects failed to exhibit any sensitivity to the direction of the causal relations. To account for these findings, I proposed that some reasoners all of the time, and all reasoners some of the time, treat a set of causal links as a "spreading activation" network in which the presence of any variable in the network increases the likely presence of any other. For example, when inferring an independent cause from the absence of an effect, the presence of the other cause will "raise the activation level" of the effect node, which in turn will raise the level of to-be predicted cause node, violating conditional independence.

It is conceivable that this tendency to reason associatively also contributed to the violations in independence observed in the present experiments. Future experiments could distinguish the UM and associative reasoning accounts by, say, assessing cross-subnetwork inferences (e.g., in Figure 6A, $p_k(IC_{A1} \mid IC_{B1})$) for which the UM account predicts independence violations but the associative reasoning account does not. In addition, an associative reasoning account would have to be developed to also account for other types of inferences assessed here (classification and feature likelihoods).

Research has established the existence of other sorts of nonnormative causal inferences, and there is little reason to expect that the same effects do not hold for category-based judgments. For example, Fernbach, Darlow, and Sloman's (2010, 2011a) assessment of reasoning with an independent cause network revealed that inferences from a cause to an effect were insensitive to the strength of other potential causes; that is, they based their judgments just on the strength of the cause known to be present and so gave conditional probability judgments that were too low (see also Fernbach et al., 2011b).[8] In addition, research reviewed by Rottman and Hastie (2014) suggests that diagnostic inferences from an effect to a cause are not sufficiently sensitive to the base rate of the cause, and the same is likely to hold when predicting category features. In summary, the claim that categories are represented as causal models is not to argue for the rationality of category-based inferences. Rather, it is to claim that many category-based inferences inherit the properties of causal cognition, often for the better but sometimes for the worse.

Implications for Models of Causal-Based Categories

Other models for representing categories' causal knowledge are unable to readily explain the present findings. As mentioned, the dependency model does not naturally represent a causal relations' functional form, treating, for example, the independent and conjunctive networks in Figures 2B and 2C as having the same dependency structure (the one in Figure 2A). Nevertheless, one might ask whether different parameterizations of that structure might account for the differences observed between the two networks—for example, perhaps independent causes yield dependency relations in Figure 2A that are stronger than conjunctive ones. However, such a proposal fails to account for the fact that the conjunctive judgments were not just weaker versions of the independent ones but rather exhibited a qualitatively different pattern. For example, conjunctive causes did not just exhibit a weaker version of the explaining away effect found with independent causes but rather an effect in the opposite direction (exoneration). The dependency model also suffers from other problems already

[8] In fact, using many of the same materials as in the present experiments, Fernbach and Rehder (2012) manipulated the strength of the links in an independent cause network and found the same thing (e.g., in Figure 2A, $p(\mathrm{IE} \mid \mathrm{IC}_1)$ was unaffected by the strength of the causal relation between IC_2 and IE).

identified in the literature. Whereas it predicts the difference in the importance of a cause and its effect (the causal status effect) should increase with causal strength, it decreases instead (Rehder & Kim, 2010, and the present Experiment 2). And, although it predicts how individual features should change in importance as a function of causal knowledge, it fails to predict any role of feature interactions and so provides no account of the different forms of coherence found here with independent and conjunctive causes.

Within the causal model framework, a linear rather than a noisy-or function for integrating the influence of independent causes was also tested but yielded an inferior fit in both experiments. Nevertheless, it should be noted that the two integration functions make predictions that are quite similar both qualitatively (the linear model yields the same patterns shown in Figures 3 and 7) and quantitatively (e.g., the correlation between their fitted predictions was $>$ .99). Indeed, whereas most subjects from Experiments 1 and 2 were fit better by the noisy-or function, the linear model yielded a better fit for 42 of 148. Thus, future research that presents test items that are better able to discriminate these accounts would be desirable (for a recent example from the domain of learning, see Cheng, Liljeholm, & Sandhofer, 2013). Of course, such tests might reveal that the integration function that best describes human judgments is a compromise between linear and noisy-or (and/or that that function varies stably over individuals).

Conclusion

The functional form of the causal relations that link features of categories affect the category-based judgments of classification, conditional feature predictions, and feature likelihoods. In addition, subjects' inferences were consistent with a noisy-or representation of independent and conjunctive causes, with the exception of a typicality effect in which typical category features imply the presence of still more typical features, an effect accounted for by the additional assumption of a UM that relates all category features. Judgments of category membership may be more susceptible to typicality effects than other judgments. Other models of causal-based categories were unable to account for these effects.

References

Ahn, W., & Kim, N. S. (2001). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 40, pp. 23–65). San Diego, CA: Academic Press.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41,* 361–416. doi:10.1006/cogp.2000.0741

Ahn, W., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory based correlations on typicality judgments. *Memory & Cognition, 30,* 107–118. doi:10.3758/BF03195270

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Barrett, S. E., Abdi, H., Murphy, G. L., & Gallagher, J. M. (1993). Theory-based correlations and their role in children's concepts. *Child Development, 64,* 1595–1616. doi:10.2307/1131458

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference*. New York, NY: Springer-Verlag. doi:10.1007/978-1-4757-2917-7

Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inference. *Journal of Memory and Language, 40,* 347–373. doi:10.1006/jmla.1998.2622

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.104.2.367

Cheng, P., Liljeholm, M., & Sandhofer, C. M. (2013). Logical consistency and objectivity in causal learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2034–2039). Austin, TX: Cognitive Science Society.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40,* 83–120. doi:10.1016/0010-0277(91)90047-8

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition, 23,* 646–658. doi:10.3758/BF03197265

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003a). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology, 15,* 161–176. doi:10.1080/09541440244000058

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003b). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31,* 581–595. doi:10.3758/BF03196099

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking and Reasoning, 11,* 349–381. doi:10.1080/13546780442000222

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21,* 329–336. doi:10.1177/0956797610361430

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011a). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140,* 168–185. doi:10.1037/a0022100

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011b). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition, 119,* 459–467. doi:10.1016/j.cognition.2011.01.013

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1327–1343. doi:10.1037/a0031851

Fernbach, P., & Rehder, B. (2012). Toward an effort reduction framework for causal inference. *Argument and Computation, 4,* 1–25.

Frosch, C. A., & Byrne, R. M. J. (2012). Causal conditionals and counterfactuals. *Acta Psychologica, 141,* 54–66. doi:10.1016/j.actpsy.2012.07.001

Geiger, S., & Oberauer, K. (2007). Reasoning with conditionals? Does every counterexample count? It's frequency that counts. *Memory & Cognition, 35,* 2060–2074. doi:10.3758/BF03192938

Gelman, S. A. (2003). *The essential child: The origins of essentialism in everyday thought*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195154061.001.0001

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195. doi:10.1016/0749-596X(88)90072-1

Hayes, B. K., & Rehder, B. (2012). Causal categorization in children and adults. *Cognitive Science, 36,* 1102–1128. doi:10.1111/j.1551-6709.2012.01244.x

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3,* 1–24. doi:10.1016/0022-1031(67)90034-0

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34,* 169–188. doi:10.1017/S0140525X10003134

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151–174). Morristown, NJ: General Learning Press.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science, 34,* 1185–1243. doi:10.1111/j.1551-6709.2010.01128.x

Khemlani, S. S., & Oppenheimer, D. M. (2011). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological Bulletin, 137,* 195–210. doi:10.1037/a0021809

Kim, N. S., & Ahn, W. (2002a). Clinical psychologists' theory-based representations of mental disorders affect their diagnostic reasoning and memory. *Journal of Experimental Psychology: General, 131,* 451–476. doi:10.1037/0096-3445.131.4.451

Kim, N. S., & Ahn, N. S. (2002b). The influence of naive causal theories on lay concepts of mental illness. *American Journal of Psychology, 115,* 33–65. doi:10.2307/1423673

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31,* 155–165. doi:10.3758/BF03196090

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34,* 113–147. doi:10.1111/j.1551-6709.2009.01058.x

Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory & Cognition, 29,* 736–744. doi:10.3758/BF03200476

Marsh, J., & Ahn, W. (2006). The role of causal status versus inter-feature links in feature weighting. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 561–566). Mahwah, NJ: Erlbaum.

Martin, J. B., & Rehder, B. (2013). *Causal knowledge and information search during categorization*. Paper presented at the 46th annual meeting of the Society for Mathematical Psychology. Potsdam, Germany.

Mayrhofer, R., Hagmayer, Y., & Waldmann, M. R. (2010). *Agents and causes: A Bayesian error attribution model of causal reasoning*. Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society. Austin, TX: The Cognitive Science Society.

Mayrhofer, R., & Rothe, A. (2012). Causal status meets coherence: The explanatory role of causal models in categorization. *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 743–748). Austin, TX: Cognitive Science Society.

Mayrhofer, R., & Waldmann, M. R. (2013). *Agents and causes: Dispositional intuitions as a guide to causal structure*. Unpublished manuscript.

McClure, J. (1998). Discounting causes of behavior: Are two reasons better than one? *Journal of Personality and Social Psychology, 74,* 7–20. doi:10.1037/0022-3514.74.1.7

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9780511529863.009

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102,* 331–355. doi:10.1037/0033-295X.102.2.331

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberchien (Ed.), *Advances in cognitive science: Vol. 2. Theory and applications* (pp. 23–45). Chichester, England: Ellis Horwood.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455–485. doi:10.1037/0033-295X.111.2.455

Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science, 15,* 100–105. doi:10.1111/j.0963-7214.2004.01502005.x

Oppenheimer, D. M., Tenenbaum, J. B., & Krynski, T. R. (2013). Categorization as causal explanation: Discounting and augmenting in a Bayesian framework. *Psychology of Learning and Motivation, 58,* 203–231. doi:10.1016/B978-0-12-407237-4.00006-2

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67,* 186–216. doi:10.1016/j.cogpsych.2013.09.002

Pearce, J. M. (1987). A model of stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61–73. doi:10.1037/0033-295X.94.1.61

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587–607. doi:10.1037/0033-295X.101.4.587

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior, 30,* 73–95. doi:10.3758/BF03192911

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufman.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, United Kingdom: Cambridge University Press.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science, 27,* 709–748. doi:10.1207/s15516709cog2705_2

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1141–1159. doi:10.1037/0278-7393.29.6.1141

Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory & Cognition, 34,* 3–16. doi:10.3758/BF03193382

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33,* 301–344. doi:10.1111/j.1551-6709.2009.01015.x

Rehder, B. (2010). Causal-based classification: A review. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 52, pp. 39–116). San Diego, CA: Elsevier Academic Press. doi:10.1016/S0079-7421(10)52002-4

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72,* 54–107. doi:10.1016/j.cogpsych.2014.02.002

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50,* 264–314. doi:10.1016/j.cogpsych.2004.09.002

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130,* 323–360. doi:10.1037/0096-3445.130.3.323

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 659–683. doi:10.1037/0278-7393.32.4.659

Rehder, B., & Kim, S. (2009). Classification as diagnostic reasoning. *Memory & Cognition, 37,* 715–729. doi:10.3758/MC.37.6.715

Rehder, B., & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1171–1206. doi:10.1037/a0019765

Rehder, B., & Ross, B. H. (2001). Abstract coherent concepts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1261–1275. doi:10.1037/0278-7393.27.5.1261

Rottman, B., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140,* 109–139. doi:10.1037/a0031903

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189–228. doi:10.1207/s15516709cog2202_2

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303–333.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. New York, NY: Springer-Verlag.

Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. *Canadian Journal of Experimental Psychology, 49,* 1–60. doi:10.1037/1196-1961.49.1.1

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning* (pp. 301–336). New York, NY: Academic Press.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31,* 233–256.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124,* 181–206. doi:10.1037/0096-3445.124.2.181

Walsh, C. R., & Sloman, S. A. (2004). Revising causal beliefs. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1423–1429). Mahwah, NJ: Lawrence Erlbaum Associates.

Walsh, C. R., & Sloman, S. A. (2008). Updating beliefs with causal models: Violations of screening off. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A Festschrift for Gordon Bower* (pp. 345–358). New York, NY: Taylor & Francis.

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 449–468. doi:10.1037/0278-7393.21.2.449

Zelazo, P. R., & Shultz, T. R. (1989). Concepts of potency and resistance in causal prediction. *Child Development, 60,* 1307–1315. doi:10.2307/1130922

Appendix

Underlying Mechanism Model

The joint distribution for the underlying mechanism (UM) model in Figure 2B is,

$$p_k(IC_1, IC_2, IE, CC_1, CC_2, CE, UM) = p_k(IE \mid IC_1, IC_2, UM)$$
$$p_k(CE \mid CC_1, CC_2, UM)p_k(IC_1 \mid UM)$$
$$p_k(IC_2 \mid UM)p_k(CC_1 \mid UM)p_k(CC_2 \mid UM)p_k(UM).$$

Assuming that the UM's causal influence on a category feature (m_{UM}) combines with its other influences according to a noisy-or function yields,

$$p_k(IE = 1 \mid IC_1, IC_2, UM) = 1 - (1 - b_{IE}) \times (1 - m_{UM})^{ind(UM)} \prod_{i=1,2} (1 - m_{IC_i,IE})^{ind(IC_i)}$$

$$p_k(CE = 1 \mid CC_1, CC_2) = 1 - (1 - b_{CE})(1 - m_{UM})^{ind(UM)} \times (1 - m_{CC_1,CC_2,CE})^{ind(CC_1)ind(CC_2)}$$

$$p_k(IC_i = 1 \mid UM) = 1 - (1 - c_{IC_i})(1 - m_{UM})^{ind(UM)}$$

$$p_k(CC_i = 1 \mid UM) = 1 - (1 - c_{CC_i})(1 - m_{UM})^{ind(UM)}.$$

The joint distribution for the observable features (IC_1, IC_2, IE, CC_1, CC_2, and CE) can be derived by a weighted sum over the two possible states of the UM,

$$p_k(IC_1, IC_2, IE, CC_1, CC_2, CE) = p_k(IC_1, IC_2, IE, CC_1, CC_2, CE, UM = 1)c_{UM} + p_k(IC_1, IC_2, IE, CC_1, CC_2, CE, UM = 0)(1 - c_{UM}).$$

Received September 16, 2013
Revision received June 5, 2014
Accepted June 9, 2014 ■